ady_ohad_ofir.ipynb

AUTHOR
Yaniv Erlich

DATE
May 12, 2021

Submitters:

Ady Stein (303791297), Ohad Edelstein (039065313), Ofir Nesher (204502926)

Yaniv Erlich
11:23 PM, May 12

Ady Ohad and Ofir: A really great and thorough job. You clearly understand the material and it seems that you had a lot of fun with this project. Excellent job. Grade: 100/100.

Exercise 1: Designing an mRNA vaccine against the south African variant

https://docs.google.com/document/d/1REzNz3VHjzsZ2iQVMg57RvN_nYjc5prMMER2h5eaCoM/edit

Hello! You were recruited by the Pfizer/BioNTech Bioinformatics team. As you know, the vaccine operation is going great but we are concerned that the vaccine is not working well for the South African variant. Your mission is to update the mRNA vaccine based on the South African Spike!

Prerequisites

Get Started 5

Install the BioPython package.

```
!pip install biopython
```

```
Requirement already satisfied: biopython in /usr/local/lib/python3.7/dist-packages (1.78)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from biopython) (1.19.5)
```

```
!pip install iso3166
```

```
Requirement already satisfied: iso3166 in /usr/local/lib/python3.7/dist-packages (1.0.1)
```

```
from collections import Counter, defaultdict
import re
from datetime import datetime

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import lxml.etree as ET

from Bio import Entrez, SeqIO, Seq
from Bio.SeqUtils import GC
from Bio import Seq

from iso3166 import countries
```

Step 2

Download the mRNA design of the Pfizer and Moderna vaccines from: https://github.com/NAalytics/Assemblies-of-putative-SARS-CoV2-spike-encoding-mRNA-sequences-for-vaccines-BNT-162b2-and-mRNA-1273

```
!git clone https://github.com/NAalytics/Assemblies-of-putative-SARS-CoV2-spike-encoding-mRNA-sequences-for-vaccines-BNT-162b2-a
```

```
fatal: destination path 'Assemblies-of-putative-SARS-CoV2-spike-encoding-mRNA-sequences-for-vaccines-BNT-162b2-and-mRNA-1273' al
```

```
!ls Assemblies-of-putative-SARS-CoV2-spike-encoding-mRNA-sequences-for-vaccines-BNT-162b2-and-mRNA-1273
```

```
'Assemblies of putative SARS-CoV2-spike-encoding mRNA sequences for vaccines BNT-162b2 and mRNA-1273.docx.pdf'
 Figure1Figure2_032321.fasta
```

```
README.md
```

```
mRNA_pfizer_moderna_vaccine_designs_filepath = \
    'Assemblies-of-putative-SARS-CoV2-spike-encoding-mRNA-sequences-for-vaccines-BNT-162b2-and-mRNA-1273/Figure1Figure2_032321.f
```

```
for record in SeqIO.parse(mRNA_pfizer_moderna_vaccine_designs_filepath, "fasta"):
    print(record)
    print('---------')
```

```
ID: Figure1_032321_Spike-encoding_contig_assembled_from_BioNTech/Pfizer_BNT-162b2_vaccine
Name: Figure1_032321_Spike-encoding_contig_assembled_from_BioNTech/Pfizer_BNT-162b2_vaccine
Description: Figure1_032321_Spike-encoding_contig_assembled_from_BioNTech/Pfizer_BNT-162b2_vaccine
Number of features: 0
Seq('GAGAATAAACTAGTATTCTTCTGGTCCCCACAGACTCAGAGAGAACCCGCCACC...GCA')
---------
ID: Figure_2_32321_Spike-encoding_contig_assembled_from_Moderna_mRNA-1273_vaccine
Name: Figure_2_32321_Spike-encoding_contig_assembled_from_Moderna_mRNA-1273_vaccine
Description: Figure_2_32321_Spike-encoding_contig_assembled_from_Moderna_mRNA-1273_vaccine
Number of features: 0
Seq('GGGAAATAAGAGAGAAAAGAAGAGTAAGAAGAAATATAAGACCCCGGCGCCGCC...AAA')
---------
```

```
vaccines_mRNA_seq = {}

for record in SeqIO.parse(mRNA_pfizer_moderna_vaccine_designs_filepath, "fasta"):
    # The sequences are actually DNA, so we need to first transcribe them to get the RNA seq.
    vaccines_mRNA_seq[re.search('Moderna|BioNTech[^_]*', record.id)[0]] = record.seq.transcribe()

vaccines_mRNA_seq
```

```
{'BioNTech/Pfizer': Seq('GAGAAUAAACUAGUAUUCUUCUGGUCCCCACAGACUCAGAGAGAACCCGCCACC...GCA'),
 'Moderna': Seq('GGGAAAUAAGAGAGAAAAGAAGAGUAAGAAGAAAUAUAAGACCCCGGCGCCGCC...AAA')}
```

```
len(vaccines_mRNA_seq['BioNTech/Pfizer']), len(vaccines_mRNA_seq['Moderna'])
```

```
(4175, 4004)
```

Step 3

Download the sequence of “Wuhan” SARS-CoV-2 sequence from NCBI Genbank accession number NC_045512.2

```
def get_genome(id):
    # this function downloads a genome from the NBCI database (seen in lecture 2)
    handle = Entrez.efetch(db='nucleotide', id=id, rettype='gb', retmode='text', email='no@one.com')
```

```
    genome = SeqIO.read(handle, 'genbank')

    return genome


wuhan_sars_cov_2_sequence = get_genome('NC_045512.2')


print(wuhan_sars_cov_2_sequence.name)
print(wuhan_sars_cov_2_sequence.description)
# print(wuhan_sars_cov_2_sequence.seq)
```

```
NC_045512
Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
```

Step 4

The NCBI Virus database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) has over 85,000 complete sequences of SARS-CoV-2. Download all complete sequences of SARS-CoV-2 in a FASTA format (You don't need these sequences until question #15)

Used the following page to search and find (complete) variants, and (manually) download them in FASTA format, to be uploaded to gDrive:

https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Severe%20acute%20respiratory%20syndrome%20coronavirus%202,%20taxid:2697049&Completeness_s=complete&HostLineage_ss=Homo%20sapiens%20(human),%20taxid:9606

Questions

Basics:

Yaniv Erlich
11:02 PM, May 12

Very nice. Correct.

1

Create a function that takes mRNA as an input and returns the translated coding region (Hint 1: you need to find the initiation codon and the stop codon. Hint 2: use BioPython!).

```
def get_translated_region(mRNA_seq, start_codons=['AUG'], end_codons=['UGA', 'UAA', 'UAG'], include_start_codon=True, include_e
    if not isinstance(mRNA_seq, str):
        mRNA_seq = str(mRNA_seq)

    coding_region_match = re.search(rf'(?:.{{3}})*?((?:{"|".join(start_codons)})(?:.{{3}})*?(?:{"|".join(end_codons)}))', mRNA_

    if coding_region_match is None:
        raise IndexError('Either the start codon or any of the end codons were not found in the mRNA sequence.')

    translated_seq = Seq.Seq(coding_region_match[1]).translate()
    if not include_start_codon:
        translated_seq = translated_seq[1:]
    if not include_end_codon:
        translated_seq = translated_seq[:-1]

    start_idx = re.search(rf'(?:.{{3}})*?((?:{"|".join(start_codons)}))', mRNA_seq).span()[1] - 3
    end_idx = start_idx + len(coding_region_match[1]) - 3

    actual_end_codon = coding_region_match[1][-3:]

    return {
        'Amino Seq': translated_seq,
        'Nucleotide Seq': coding_region_match[1],
        'End Codon': actual_end_codon,
        'Start Codon Index': start_idx,
        'End Codon Index': end_idx
    }
```

```
vaccines_data = {company_name: get_translated_region(vaccines_mRNA_seq[company_name]) for company_name in vaccines_mRNA_seq}

for company_name in vaccines_data:
    print(f'{company_name}:\n======================')
    for key in vaccines_data[company_name]:
        if 'Seq' not in key:
            print(f'- {key}: {vaccines_data[company_name][key]}')
    print(f'- Length of the coding sequence is {len(vaccines_data[company_name]["Nucleotide Seq"]) - 3} nucleotides ({len(vacci
    print()
```

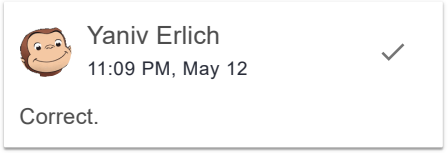
```
BioNTech/Pfizer:
=======================
- End Codon: UGA
- Start Codon Index: 54
- End Codon Index: 3873
- Length of the coding sequence is 3819 nucleotides (1273 amino acids), including the start codon but excluding the end codon.

Moderna:
=======================
- End Codon: UGA
- Start Codon Index: 57
- End Codon Index: 3876
- Length of the coding sequence is 3819 nucleotides (1273 amino acids), including the start codon but excluding the end codon.
```

2

For each vaccine:

1. What's the position of the initiation codon ("AUG")? (All positions should be reported as zero based)
2. What's the position of the termination codon?
3. What sequence did each company use for the stop codon?
4. Please suggest another stop codon sequence.
5. How would you call the area between the first nucleotide and the "AUG" site?
6. How would you call the area between the termination codon and the last nucleotide?



BioNTech/Pfizer:

1. Start Codon Index: 54 (0 based)
2. End Codon Index: 3876
3. End Codon: UGA

Moderna:

1. Start Codon Index: 57 (0 based)
2. End Codon Index: 3879
3. End Codon: UGA

Both:

4. Other possible end codon sequences are: **UAA**, **UAG**.
5. The area between the first nucleotide and the **AUG** site is called the 5' UTR (untraslated region).
6. The area between the termination codon and the last nucleotide is called the 3' UTR (untraslated region).

3

Using your function, translate the coding area into protein. What's the length of each coding sequence?

```
# for company_name in vaccines_data:
#     print(company_name)
#     print(f'Translates coding area into protein: {vaccines_data[company_name]["Amino Seq"]}:\n======================')
```

Yaniv Erlich
11:09 PM, May 12

Correct.

BioNTech/Pfizer:

- Length of the coding sequence is 3819 nucleotides (1273 amino acids), including the start codon but excluding the end codon.

Moderna:

- Length of the coding sequence is 3819 nucleotides (1273 amino acids), including the start codon but excluding the end codon.

4

Are the coding regions of the vaccines identical in the protein level?

```
company_names = list(vaccines_data.keys())
company_names
```

```
['BioNTech/Pfizer', 'Moderna']
```

```
vaccines_data[company_names[0]]['Amino Seq'] == vaccines_data[company_names[1]]['Amino Seq']
```

```
True
```

Yaniv Erlich
11:09 PM, May 12
Correct

The coding regions of the vaccines **are identical** on the **protein** level.

5

Are the coding regions of the vaccines identical in the nucleotide level?

Yaniv Erlich
11:09 PM, May 12
Correct

```
vaccines_data[company_names[0]]['Nucleotide Seq'] == vaccines_data[company_names[1]]['Nucleotide Seq']
```

```
False
```

The coding regions of the vaccines are **NOT identical** on the **nucleotide** level.

##**Codon usage**:

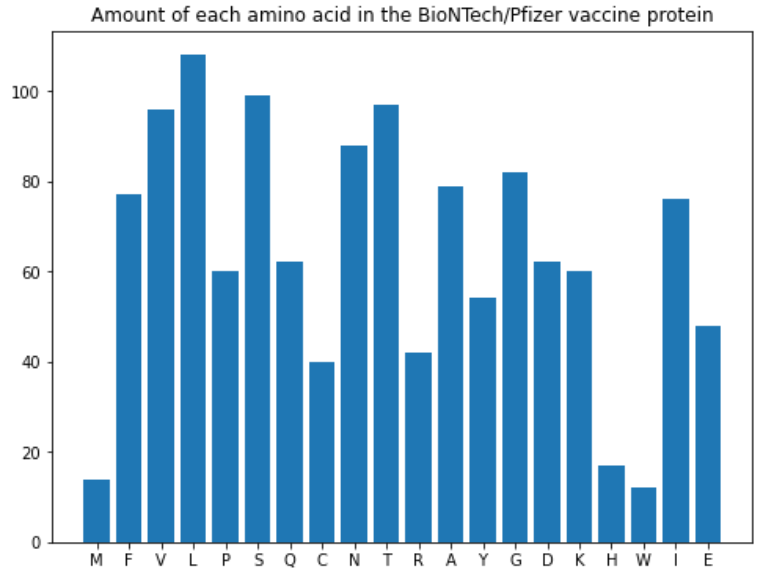
6

Create a histogram of amino-acids usage for the Pfizer vaccine

```
amino_counter = Counter(str(vaccines_data[company_names[0]]['Amino Seq']))

plt.figure(figsize=(8, 6))
plt.bar(amino_counter.keys(), amino_counter.values())
plt.title(f'Amount of each amino acid in the {company_names[0]} vaccine protein');
```

Yaniv Erlich
11:09 PM, May 12
Correct



7

Create the same histogram but color each codon within the amino-acid (i.e. create a stacked histogram)

```
nucleotide_seq = str(vaccines_data[company_names[0]]['Nucleotide Seq'])

amino_nucleotide_counter = defaultdict(lambda: defaultdict(int))
for amino_idx, amino in enumerate(str(vaccines_data[company_names[0]]['Amino Seq'])):
    for codon_sub_idx in range(3):
        amino_nucleotide_counter[amino][nucleotide_seq[(amino_idx * 3) + codon_sub_idx]] += 1

amino_nucleotide_counter
```

```
defaultdict(<function __main__.<lambda>>,
            {'A': defaultdict(int, {'A': 4, 'C': 144, 'G': 79, 'U': 10}),
             'C': defaultdict(int, {'C': 27, 'G': 40, 'U': 53}),
             'D': defaultdict(int, {'A': 62, 'C': 44, 'G': 62, 'U': 18}),
             'E': defaultdict(int, {'A': 62, 'G': 82}),
             'F': defaultdict(int, {'C': 62, 'U': 169}),
             'G': defaultdict(int, {'A': 12, 'C': 66, 'G': 167, 'U': 1}),
             'H': defaultdict(int, {'A': 17, 'C': 32, 'U': 2}),
             'I': defaultdict(int, {'A': 76, 'C': 66, 'U': 86}),
             'K': defaultdict(int, {'A': 129, 'G': 51}),
             'L': defaultdict(int, {'C': 111, 'G': 105, 'U': 108}),
             'M': defaultdict(int, {'A': 14, 'G': 14, 'U': 14}),
             'N': defaultdict(int, {'A': 176, 'C': 67, 'U': 21}),
             'P': defaultdict(int, {'A': 4, 'C': 146, 'U': 30}),
             'Q': defaultdict(int, {'A': 66, 'C': 62, 'G': 58}),
```

```
            'R': defaultdict(int, {'A': 43, 'C': 21, 'G': 62}),
            'S': defaultdict(int, {'A': 64, 'C': 121, 'G': 64, 'U': 48}),
            'T': defaultdict(int, {'A': 131, 'C': 160}),
            'V': defaultdict(int, {'C': 8, 'G': 183, 'U': 97}),
            'W': defaultdict(int, {'G': 24, 'U': 12}),
            'Y': defaultdict(int, {'A': 54, 'C': 49, 'U': 59})})
```

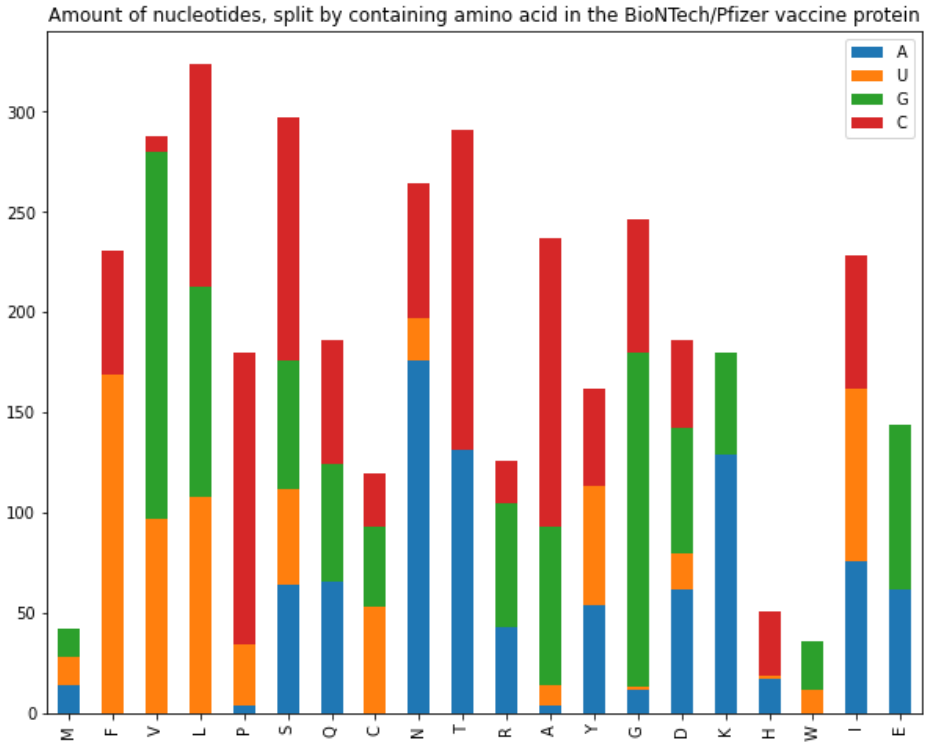
```
df = pd.DataFrame(amino_nucleotide_counter)
df
```

	M	F	V	L	P	S	Q	C	N	T	R	A	Y	G	D	K	H	W	I	E
A	14.0	NaN	NaN	NaN	4.0	64	66.0	NaN	176.0	131.0	43.0	4	54.0	12	62	129.0	17.0	NaN	76.0	62.0
U	14.0	169.0	97.0	108.0	30.0	48	NaN	53.0	21.0	NaN	NaN	10	59.0	1	18	NaN	2.0	12.0	86.0	NaN
G	14.0	NaN	183.0	105.0	NaN	64	58.0	40.0	NaN	NaN	62.0	79	NaN	167	62	51.0	NaN	24.0	NaN	82.0
C	NaN	62.0	8.0	111.0	146.0	121	62.0	27.0	67.0	160.0	21.0	144	49.0	66	44	NaN	32.0	NaN	66.0	NaN

Yaniv Erlich
11:10 PM, May 12

Hmm... That was not the question. You created a histogram of the nucleotides but the question was about *codons*
-5pt

```
df.transpose().plot(kind="bar", stacked=True, figsize=(10, 8))
plt.title(f'Amount of nucleotides, split by containing amino acid in the {company_names[0]} vaccine protein');
```



Comparison to the Wuhan strain:

8

What's the length of the Wuhan strain genome?

```
print(f'The length of the Wuhan strain genome is {len(wuhan_sars_cov_2_sequence.seq)} nucleotides.')
```

```
The length of the Wuhan strain genome is 29903 nucleotides.
```

Yaniv Erlich
11:10 PM, May 12

Correct

9

Report the nucleotide sequence of the Spike region of the Wuhan strain from the initiation codon to the termination codon (hint: use the information in the NCBI website about the starting and stopping position of Spike)

```
# From https://www.ncbi.nlm.nih.gov/gene/43740568
wuhan_sars_cov_2_data = get_translated_region(str(wuhan_sars_cov_2_sequence.seq[21563 - 1: 25384].transcribe()))
```

```
# Number of characters to print per row
step = 102

for start_idx in range(0, len(wuhan_sars_cov_2_data['Nucleotide Seq']), step):
    print(wuhan_sars_cov_2_data['Nucleotide Seq'][start_idx:start_idx + step])
```

```
AUGUUUGUUUUUCUUGUUUUAUUGCCACUAGUCUCUAGUCAGUGUGUUAAUCUUACAACCAGAACUCAAUUACCCCCUGCAUACACUAAUUCUUUCACACGU
GGUGUUUAUUACCCUGACAAAGUUUUCAGAUCCUCAGUUUUACAUUCAACUCAGGACUUGUUCUUACCUUUCUUUUCCAAUGUUACUUGGUUCCAUGCUAUA
CAUGUCUCUGGGACCAAUGGUACUAAGAGGUUUGAUAACCCUGUCCUACCAUUUAAUGAUGGUGUUUAUUUUGCUUCCACUGAGAAGUCUAACAUAAUAAGA
GGCUGGAUUUUUGGUACUACUUUAGAUUCGAAGACCCAGUCCCUACUUAUUGUUAAUAACGCUACUAAUGUUGUUAUUAAAGUCUGUGAAUUUCAAUUUUGU
AAUGAUCCAUUUUUGGGUGUUUAUUACCACAAAAACAACAAAAGUUGGAUGGAAAGUGAGUUCAGAGUUUAUUCUAGUGCGAAUAAUUGCACUUUUGAAUAU
GUCUCUCAGCCUUUUCUUAUGGACCUUGAAGGAAAACAGGGUAAUUUCAAAAAUCUUAGGGAAUUUGUGUUUAAGAAUAUUGAUGGUUAUUUUAAAAUAUAU
UCUAAGCACACGCCUAUUAAUUUAGUGCGUGAUCUCCCUCAGGGUUUUUCGGCUUUAGAACCAUUGGUAGAUUUGCCAAUAGGUAUUAACAUCACUAGGUUU
CAAACUUUACUUGCUUUACAUAGAAGUUAUUUGACUCCUGGUGAUUCUUCUUCAGGUUGGACAGCUGGUGCUGCAGCUUAUUAUGUGGGUUAUCUUCAACCU
```


Yaniv Erlich
11:12 PM, May 12

Correct. As a side comment: it is a bad practice to encode positions as constants and not variables. You would look at this code in one year, you will not even remember what it means. So a better solution is to say:

spike_start = 21563
spike_end = 25384
and use those variables to get the sequence (edited)

```
AGGACUUUUCUAUUAAAAUAUAAUGAAAAUGGAACCAUUACAGAUGCUGUAGACUGUGCACUUGACCCUCUCUCAGAAACAAAGUGUACGUUGAAAUCCUUC
ACUGUAGAAAAAGGAAUCUAUCAAACUUCUAACUUUAGAGUCCAACCAACAGAAUCUAUUGUUAGAUUUCCUAAUAUUACAAACUUGUGCCCUUUUGGUGAA
GUUUUUAACGCCACCAGAUUUGCAUCUGUUUAUGCUUGGAACAGGAAGAGAAUCAGCAACUGUGUUGCUGAUUAUUCUGUCCUAUAUAAUUCCGCAUCAUUU
UCCACUUUUAAGUGUUAUGGAGUGUCUCCUACUAAAUUAAAUGAUCUCUGCUUUACUAAUGUCUAUGCAGAUUCAUUUGUAAUUAGAGGUGAUGAAGUCAGA
CAAAUCGCUCCAGGGCAAACUGGAAAGAUUGCUGAUUAUAAUUAUAAAUUACCAGAUGAUUUUACAGGCUGCGUUAUAGCUUGGAAUUCUAACAAUCUUGAU
UCUAAGGUUGGUGGUAAUUAUAAUUACCUGUAUAGAUUGUUUAGGAAGUCUAAUCUCAAACCUUUUGAGAGAGAUAUUUCAACUGAAAUCUAUCAGGCCGGU
AGCACACCUUGUAAUGGUGUUGAAGGUUUUAAUUGUUACUUUCCUUUACAAUCAUAUGGUUUCCAACCCACUAAUGGUGUUGGUUACCAACCAUACAGAGUA
GUAGUACUUUCUUUUGAACUUCUACAUGCACCAGCAACUGUUUGUGGACCUAAAAAGUCUACUAAUUUGGUUAAAAACAAAUGUGUCAAUUUCAACUUCAAU
GGUUUAACAGGCACAGGUGUUCUUACUGAGUCUAACAAAAAGUUUCUGCCUUUCCAACAAUUUGGCAGAGACAUUGCUGACACUACUGAUGCUGUCCGUGAU
CCACAGACACUUGAGAUUCUUGACAUUACACCAUGUUCUUUUGGUGGUGUCAGUGUUAUAACACCAGGAACAAAUACUUCUAACCAGGUUGCUGUUCUUUAU
CAGGAUGUUAACUGCACAGAAGUCCCUGUUGCUAUUCAUGCAGAUCAACUUACUCCUACUUGGCGUGUUUAUUCUACAGGUUCUAAUGUUUUUCAAACACGU
GCAGGCUGUUUAAUAGGGGCUGAACAUGUCAACAACUCAUAUGAGUGUGACAUACCCAUUGGUGCAGGUAUAUGCGCUAGUUAUCAGACUCAGACUAAUUCU
CCUCGGCGGGCACGUAGUGUAGCUAGUCAAUCCAUCAUUGCCUACACUAUGUCACUUGGUGCAGAAAAUUCAGUUGCUUACUCUAAUAACUCUAUUGCCAUA
CCCACAAAUUUUACUAUUAGUGUUACCACAGAAAUUCUACCAGUGUCUAUGACCAAGACAUCAGUAGAUUGUACAAUGUACAUUUGUGGUGAUUCAACUGAA
UGCAGCAAUCUUUUGUUGCAAUAUGGCAGUUUUUGUACACAAUUAAACCGUGCUUUAACUGGAAUAGCUGUUGAACAAGACAAAAACACCCAAGAAGUUUUU
GCACAAGUCAAACAAAUUUACAAAACACCACCAAUUAAAGAUUUUGGUGGUUUUAAUUUUUCACAAAUAUUACCAGAUCCAUCAAAACCAAGCAAGAGGUCA
UUUAUUGAAGAUCUACUUUUCAACAAAGUGACACUUGCAGAUGCUGGCUUCAUCAAACAAUAUGGUGAUUGCCUUGGUGAUAUUGCUGCUAGAGACCUCAUU
UGUGCACAAAAGUUUAACGGCCUUACUGUUUUGCCACCUUUGCUCACAGAUGAAAUGAUUGCUCAAUACACUUCUGCACUGUUAGCGGGUACAAUCACUUCU
GGUUGGACCUUUGGUGCAGGUGCUGCAUUACAAAUACCAUUUGCUAUGCAAAUGGCUUAUAGGUUUAAUGGUAUUGGAGUUACACAGAAUGUUCUCUAUGAG
AACCAAAAAUUGAUUGCCAACCAAUUUAAUAGUGCUAUUGGCAAAAUUCAAGACUCACUUUCUUCCACAGCAAGUGCACUUGGAAAACUUCAAGAUGUGGUC
AACCAAAAUGCACAAGCUUUAAACACGCUUGUUAAACAACUUAGCUCCAAUUUUGGUGCAAUUUCAAGUGUUUUAAAUGAUAUCCUUUCACGUCUUGACAAA
GUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCAGAAAUCAGAGCU
UCUGCUAAUCUUGCUGCUACUAAAAUGUCAGAGUGUGUACUUGGACAAUCAAAAAGAGUUGAUUUUUGUGGAAAGGGCUAUCAUCUUAUGUCCUUCCCUCAG
UCAGCACCUCAUGGUGUAGUCUUCUUGCAUGUGACUUAUGUCCCUGCACAAGAAAAGAACUUCACAACUGCUCCUGCCAUUUGUCAUGAUGGAAAAGCACAC
UUUCCUCGUGAAGGUGUCUUUGUUUCAAAUGGCACACACUGGUUUGUAACACAAAGGAAUUUUUAUGAACCACAAAUCAUUACUACAGACAACACAUUUGUG
UCUGGUAACUGUGAUGUUGUAAUAGGAAUUGUCAACAACACAGUUUAUGAUCCUUUGCAACCUGAAUUAGACUCAUUCAAGGAGGAGUUAGAUAAAUAUUUU
AAGAAUCAUACAUCACCAGAUGUUGAUUUAGGUGACAUCUCUGGCAUUAAUGCUUCAGUUGUAAACAUUCAAAAAGAAAUUGACCGCCUCAAUGAGGUUGCC
AAGAAUUUAAAUGAAUCUCUCAUCGAUCUCCAAGAACUUGGAAAGUAUGAGCAGUAUAUAAAAUGGCCAUGGUACAUUUGGCUAGGUUUUAUAGCUGGCUUG
AUUGCCAUAGUAAUGGUGACAAUUAUGCUUUGCUGUAUGACCAGUUGCUGUAGUUGUCUCAAGGGCUGUUGUUCUUGUGGAUCCUGCUGCAAAUUUGAUGAA
GACGACUCUGAGCCAGUGCUCAAAGGAGUCAAAUUACAUUACACAUAA
```

```
len(wuhan_sars_cov_2_data['Nucleotide Seq'])
```

```
3822
```

10

Report the GC content of the Spike region of the Wuhan strain and compare it to the GC content of the Spike region of the Pfizer vaccine

```
# [Ady - Note]: Changed this on Fri morning (before 11:30) when I noticed we had a mistake here
# (we previously looked at the entire virus instead of just the spike and got a GC content of 37.97%)
f'Wuhan virus variant spike protein GC content is: {GC(wuhan_sars_cov_2_data["Nucleotide Seq"]):.2f}%.'
```

```
'Wuhan virus variant spike protein GC content is: 37.31%.'
```

```
f'{company_names[0]} vaccine (spike protein) GC content is: {GC(vaccines_data[company_names[0]]["Nucleotide Seq"]):.2f}%.'
```

```
'BioNTech/Pfizer vaccine (spike protein) GC content is: 56.99%.'
```

Yaniv Erlich
11:13 PM, May 12

Correct.

11

Translate the sequence into amino acids and report the sequence of the protein

```
step = 68
for start_idx in range(0, len(wuhan_sars_cov_2_data['Amino Seq']), step):
    print(wuhan_sars_cov_2_data['Amino Seq'][start_idx:start_idx + step])
```

```
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI
HVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY
SKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE
VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR
QIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
STPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY
QDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
PRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS
GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF
KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

12

Compare between the Spike region of the Pfizer vaccine to the Spike of the Wuhan strain in the protein level.

Are there any differences and where?

Bonus: explain the differences!

```
print('Vac: Pfizer vaccine  -  Vir: Wuhan variant virus\n==========================================================================\n')
step = 68
for start_idx in range(0, len(vaccines_data[company_names[0]]['Amino Seq']), step):
    vac_part = vaccines_data[company_names[0]]['Amino Seq'][start_idx:start_idx + step]
    vir_part = wuhan_sars_cov_2_data['Amino Seq'][start_idx:start_idx + step]

    print('Vac:', vac_part)
    print('Vir:', vir_part)
    print('    ', "".join([' ' if vac_part[i] == vir_part[i] else '^' for i in range(len(vac_part))]))
```

Yaniv Erlich
11:13 PM, May 12

Correct and a nice presentation.

```
Vac: Pfizer vaccine  -  Vir: Wuhan variant virus
================================================================

Vac: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI
Vir: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI

Vac: HVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
Vir: HVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC

Vac: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY
Vir: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY

Vac: SKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP
Vir: SKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP

Vac: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE
Vir: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE

Vac: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR
Vir: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR

Vac: QIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
Vir: QIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG

Vac: STPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
Vir: STPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN

Vac: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY
Vir: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY

Vac: QDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
Vir: QDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS

Vac: PRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
Vir: PRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE

Vac: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS
Vir: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS

Vac: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS
Vir: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS

Vac: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV
Vir: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV

Vac: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
Vir: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
                                           ^^
Vac: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH
Vir: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH

Vac: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF
Vir: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF

Vac: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL
Vir: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL

Vac: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
Vir: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

The **PP** amino change in the virus is put there to keep the structural integrity of the spike protein, and not allow it to "transform" from its "attach to cell" form.

Based on: https://cen.acs.org/pharmaceuticals/vaccines/tiny-tweak-behind-COVID-19/98/i38

"... Once attached, the spike undergoes a dramatic transformation, stretching before partially turning inside out to forcefully fuse with our cells. ..."

"... Unfortunately for vaccine developers, spike proteins are liable to spring from their stubby prefusion shape into their elongated postfusion form on a hair trigger. ..."

"... adding two prolines—the most rigid of the 20 amino acids—to a key joint of a vaccine's spike protein could stabilize the structure's prefusion shape. This 2P mutation worked in preclinical studies of Graham and Moderna's MERS vaccine, so they applied it to Moderna's COVID-19 vaccine. ..."

"... Other companies, including Johnson & Johnson, Novavax, and Pfizer, are hoping the 2P mutation works for their COVID-19 vaccines too. ..."

13

For the amino acids that appear in the vaccine and the Wuhan strain, create a look-up table that maps between the Wuhan strain codon to the Pfizer codon

We build something that is a bit more than a LUT: we count *how many times* each codon in the vaccine was used instead of the corresponding codon in the virus.

```
vir_to_vac_codons_map = defaultdict(lambda: defaultdict(int))
```

```
for start_idx in range(0, len(vaccines_data[company_names[0]]['Amino Seq'])):
    vac_amino = vaccines_data[company_names[0]]['Amino Seq'][start_idx]
    vir_amino = wuhan_sars_cov_2_data['Amino Seq'][start_idx]

    if vac_amino == vir_amino:
        vac_codon = vaccines_data[company_names[0]]['Nucleotide Seq'][start_idx * 3:(start_idx + 1) * 3]
        vir_codon = wuhan_sars_cov_2_data['Nucleotide Seq'][start_idx * 3:(start_idx + 1) * 3]

        vir_to_vac_codons_map[vir_codon][vac_codon] += 1
```

```
vir_to_vac_codons_map
```

```
defaultdict(<function __main__.<lambda>>,
            {'AAA': defaultdict(int, {'AAA': 6, 'AAG': 31}),
             'AAC': defaultdict(int, {'AAC': 26, 'AAU': 8}),
             'AAG': defaultdict(int, {'AAA': 3, 'AAG': 20}),
             'AAU': defaultdict(int, {'AAC': 41, 'AAU': 13}),
             'ACA': defaultdict(int, {'ACA': 13, 'ACC': 27}),
             'ACC': defaultdict(int, {'ACA': 1, 'ACC': 9}),
             'ACG': defaultdict(int, {'ACC': 3}),
             'ACU': defaultdict(int, {'ACA': 20, 'ACC': 24}),
             'AGA': defaultdict(int, {'AGA': 10, 'AGG': 1, 'CGG': 9}),
             'AGC': defaultdict(int, {'AGC': 3, 'UCC': 2}),
             'AGG': defaultdict(int, {'AGA': 3, 'CGC': 1, 'CGG': 6}),
             'AGU': defaultdict(int, {'AGC': 15, 'UCU': 2}),
             'AUA': defaultdict(int, {'AUC': 16, 'AUU': 2}),
             'AUC': defaultdict(int, {'AUC': 12, 'AUU': 2}),
             'AUG': defaultdict(int, {'AUG': 14}),
             'AUU': defaultdict(int, {'AUC': 38, 'AUU': 6}),
             'CAA': defaultdict(int, {'CAA': 4, 'CAG': 42}),
             'CAC': defaultdict(int, {'CAC': 3, 'CAU': 1}),
             'CAG': defaultdict(int, {'CAG': 16}),
             'CAU': defaultdict(int, {'CAC': 12, 'CAU': 1}),
             'CCA': defaultdict(int, {'CCC': 15, 'CCU': 10}),
             'CCC': defaultdict(int, {'CCC': 3, 'CCU': 1}),
             'CCU': defaultdict(int, {'CCA': 4, 'CCC': 8, 'CCU': 17}),
             'CGC': defaultdict(int, {'CGG': 1}),
             'CGG': defaultdict(int, {'AGA': 1, 'CGG': 1}),
             'CGU': defaultdict(int, {'AGA': 7, 'CGG': 2}),
             'CUA': defaultdict(int, {'CUG': 9}),
             'CUC': defaultdict(int, {'CUG': 12}),
             'CUG': defaultdict(int, {'CUG': 3}),
             'CUU': defaultdict(int, {'CUG': 36}),
             'GAA': defaultdict(int, {'GAA': 12, 'GAG': 22}),
             'GAC': defaultdict(int, {'GAC': 12, 'GAU': 7}),
             'GAG': defaultdict(int, {'GAA': 2, 'GAG': 12}),
             'GAU': defaultdict(int, {'GAC': 32, 'GAU': 11}),
             'GCA': defaultdict(int, {'GCA': 2, 'GCC': 21, 'GCU': 4}),
             'GCC': defaultdict(int, {'GCC': 7, 'GCU': 1}),
             'GCG': defaultdict(int, {'GCC': 2}),
             'GCU': defaultdict(int, {'GCA': 2, 'GCC': 35, 'GCU': 5}),
             'GGA': defaultdict(int, {'GGA': 2, 'GGC': 13, 'GGG': 2}),
             'GGC': defaultdict(int, {'GGA': 3, 'GGC': 12}),
             'GGG': defaultdict(int, {'GGA': 2, 'GGC': 1}),
             'GGU': defaultdict(int,
                        {'GGA': 5, 'GGC': 40, 'GGG': 1, 'GGU': 1}),
             'GUA': defaultdict(int, {'GUG': 15}),
             'GUC': defaultdict(int, {'GUC': 1, 'GUG': 19, 'GUU': 1}),
             'GUG': defaultdict(int, {'GUC': 1, 'GUG': 12}),
```

Yaniv Erlich
11:14 PM, May 12

Correct. You also made me happy because many groups forgot to exclude the PP comparison.

```
'GUU': defaultdict(int, {'GUC': 6, 'GUG': 41}),
'UAC': defaultdict(int, {'UAC': 13, 'UAU': 1}),
'UAU': defaultdict(int, {'UAC': 36, 'UAU': 4}),
'UCA': defaultdict(int, {'AGC': 18, 'UCC': 4, 'UCU': 4}),
'UCC': defaultdict(int, {'AGC': 8, 'UCC': 4}),
'UCG': defaultdict(int, {'AGC': 1, 'UCU': 1}),
'UCU': defaultdict(int, {'AGC': 19, 'UCC': 12, 'UCU': 6}),
'UGC': defaultdict(int, {'UGC': 8, 'UGU': 4}),
'UGG': defaultdict(int, {'UGG': 12}),
'UGU': defaultdict(int, {'UGC': 19, 'UGU': 9}),
'UUA': defaultdict(int, {'CUC': 1, 'CUG': 27}),
'UUC': defaultdict(int, {'UUC': 15, 'UUU': 3}),
'UUG': defaultdict(int, {'CUC': 2, 'CUG': 18}),
'UUU': defaultdict(int, {'UUC': 47, 'UUU': 12})})
```

14

Can you deduce any rule about the look-up table that Pfizer uses internally?

Bonus: explain the differences!

Yaniv Erlich
11:15 PM, May 12

Question is correct but bonus is not. The increased GC content is for better translation efficiency and not double helix (RNA is a single strand).

It looks like most codons in the virus were replaced (in the vaccine) with compatible codons, but with more G and C nucleotides in them.

A few examples:

- 'AAA' was mapped to 'AAG' in more than 80% of the cases.
- 'AAU' was mapped to 'AAC' in more than 75% of the cases.
- 'AUU' was mapped to 'AUC' in more than 85% of the cases.
- 'CAA' was mapped to 'CAG' in more than 90% of the cases.
- 'UUG' was mapped to 'CAC' or 'CUG in 100% of the cases.

There are also cases that are less pronounced (e.g. 'ACU' to 'ACC' in only ~55%), and even cases where a mapping reduced the GC content (e.g. 'GGG' to 'GGA' in 66%), but both of these are the minority of cases (especially taking into account the codon counts).

This leads to an overall increase in GC content (as also seen above in question 10).

Main understanding of what a higher GC content does in this case:

Contrary to the fact the a higher GC content makes double-helix DNA more structurally sound, a higher GC content in this RNA sequence actually makes the virus weaker or "less fit" by making it less able to bind to host cells through the ACE2 protein.

Overall it looks like researchers are trying to balance having a more sound structure of the spike protein (using the **PP** change in the amino sequence), with making the virus more easy to defend against by the immune system using the different codons (which also leads to a higher GC content).

Based on: https://www.nature.com/articles/s41598-020-72533-2

"... Understanding the mechanism that leads to viral attenuation requires a thorough characterization of the viral sequence and of the consequences of sequence changes. There are several factors that may contribute to the efficacy of deoptimization strategies. In changing the codon pair usage, the dinucleotide frequency and the GC content are altered; mRNA secondary structure and translational kinetics are also perturbed. Further, the CpG content is changing, leading potentially to altered immunogenicity. It is likely that codon pair (or codon) deoptimization leads to reduced expression, either due to changes in transcription, mRNA stability, or translation efficiency[45,46]. Alternatively, it is possible that deoptimization may lead to perturbed cotranslational folding[47], resulting in altered protein conformation. In the case of the S protein, this may lead to decreased binding affinity for the ACE2 protein, thus affecting viral fitness. ..."

And this is for the ACE2 protein:

https://theconversation.com/what-is-the-ace2-receptor-how-is-it-connected-to-coronavirus-and-why-might-it-be-key-to-treating-covid-19-the-experts-explain-136928

The South-African Variant (B.1.351):

15

The 75,000 genomes that you have are NOT aligned to the Wuhan strain.

For example:

Notice that a nucleotide 13th in the genome Wuhan strain corresponds to nucleotide 14th in the genome in your collection. So you cannot simply use positions from the Wuhan strain to find positions in your collection.

Unfortunately, the only information that you have about the South Afircan variant is based on the Wuhan coordinates.

Namely, you know that (i) position 23011 in Wuhan strain is not “G” but “A” and (ii) position 23062 is not “A” but “T” (iii) and that position 21800 is not “A” but “C”.

Without aligning, develop an algorithm to quickly sift through the 75000 genomes and find all entries that can match the South Afircan variant (hint: mutations occur very rarely. You can assume that positions adjacent to the mutated sites are identical to the Wuhan strain).

Our algorithm will work in the following manner:

- We will look at a search area around the mutation site (as indexed by the Wuhan variant); e.g. if the mutation site is at 21800 in the Wuhan variant, we will look at the area 21700 - 21900 in the tested sample.
- In that area we will search for a sequence that includes the mutation; e.g. if we look at 5 nucleotides on each side of the mutation site for site 21800, in the Wuhan varinat we have "GTTTG**A**TAACC", so we will be searching for "GTTTG**C**TAACC" in the search area of the sample.
- We will do this for several search area sizes and several searched sequence lengths, to be able to better select the appropriate values or these.

- We will be assuming throught the above, that it is not necessary to work in tripletes of codons, as the chances for a problem to arise from these are low (we will however verify that our results make sense, by comparing between the different findings).

```
from google.colab import drive

GOOGLE_DRIVE_MOUNT_POINT = '/gdrive'
drive.mount(GOOGLE_DRIVE_MOUNT_POINT)
```

```
Drive already mounted at /gdrive; to attempt to forcibly remount, call drive.mount("/gdrive", force_remount=True).
```

```
# TODO: Change this to where the file is located your drive

COMPLETE_COVID_SEQUENCES_FILEPATH = f'{GOOGLE_DRIVE_MOUNT_POINT}/MyDrive/MSc/02 Genomics/HW1/sequences.fasta'


def get_area_around_site(seq, site_idx, side_length, substitute_for_site=None):
    ret = str(seq[site_idx - side_length:site_idx + 1 + side_length])

    if substitute_for_site is not None:
        ret = f'{ret[:side_length]}{substitute_for_site}{ret[side_length + 1:]}'

    return ret


def is_sample_sa_variant(seq, search_area_around_site_per_side, interesting_site_indexes, parts_to_search_around_interesting_sit
    for interesting_site_idx, interesting_site in enumerate(interesting_site_indexes):
        search_area = get_area_around_site(seq, interesting_site, search_area_around_site_per_side)
        if search_area.find(parts_to_search_around_interesting_sites_in_sa_variants[interesting_site_idx]) == -1:
            return False

    return True



interesting_site_indexes = [21800, 23011, 23062]
interesting_sites_sa_nt = ['C', 'A', 'T']

diff_res = {}
for required_match_length_per_site_side in [50, 20, 10, 5]:
    for search_area_around_site_per_side in [300, 200, 100]:

        parts_to_search_around_interesting_sites_in_sa_variants = \
            [get_area_around_site(wuhan_sars_cov_2_sequence.seq, interesting_site_indexes[i], required_match_length_per_site_sic
            for i in range(len(interesting_site_indexes))]

        complete_covid_sa_sequences = {}
        for seq_data in SeqIO.parse(COMPLETE_COVID_SEQUENCES_FILEPATH, "fasta"):
            if is_sample_sa_variant(seq_data.seq, search_area_around_site_per_side, interesting_site_indexes, parts_to_search_ar
                complete_covid_sa_sequences[seq_data.id] = seq_data

        diff_res[f'{required_match_length_per_site_side}__{search_area_around_site_per_side}'] = complete_covid_sa_sequences.key
```

Yaniv Erlich
11:17 PM, May 12

Yaniv Erlich
11:17 PM, May 12

Correct! Very nice.

```
        print(f'- Search area: {(2 * search_area_around_site_per_side) + 1}, searched sequence length: {(2 * required_match_len
```

```
- Search area: 601, searched sequence length: 101.
  Found 62 samples.

- Search area: 401, searched sequence length: 101.
  Found 62 samples.

- Search area: 201, searched sequence length: 101.
  Found 37 samples.

- Search area: 601, searched sequence length: 41.
  Found 62 samples.

- Search area: 401, searched sequence length: 41.
  Found 62 samples.

- Search area: 201, searched sequence length: 41.
  Found 62 samples.

- Search area: 601, searched sequence length: 21.
  Found 64 samples.

- Search area: 401, searched sequence length: 21.
  Found 64 samples.

- Search area: 201, searched sequence length: 21.
  Found 64 samples.

- Search area: 601, searched sequence length: 11.
  Found 64 samples.

- Search area: 401, searched sequence length: 11.
  Found 64 samples.

- Search area: 201, searched sequence length: 11.
  Found 64 samples.
```

```
 # Make sure that when there are less or equal than the max number of samples,
 # the samples are a subset of the set with the max samples (as we'd expect).
 for k, sample_ids in diff_res.items():
     print(f'"{k}" is a subsample of "5__300": {len(set(diff_res["5__300"]).intersection(set(sample_ids))) == len(sample_ids)}')
```

```
"50__300" is a subsample of "5__300": True
"50__200" is a subsample of "5__300": True
"50__100" is a subsample of "5__300": True
"20__300" is a subsample of "5__300": True
"20__200" is a subsample of "5__300": True
"20__100" is a subsample of "5__300": True
"10__300" is a subsample of "5__300": True
"10__200" is a subsample of "5__300": True
"10__100" is a subsample of "5__300": True
"5__300" is a subsample of "5__300": True
```

```
"5__200" is a subsample of "5__300": True
"5__100" is a subsample of "5__300": True
```

- We found there are 64 such samples, if we require to find a nucleotide sequence of length 21 (or less), within an area of 201 nucleotides (or more).
- Searching for longer sequences (especially when the search area is on the smaller side) would result in less sequences found (but a subset of the 64 sample we found, not different ones). **Note**: This means there's at least one other mutation/sequencing error between 20 and 50 nucleotides away from one of the existing mutation sites, or that there's a large "misalignment" that prevents finding the requested sequence in the search area.

Yaniv Erlich
11:17 PM, May 12

Correct.

16

Report all entries in your collection that you think match the South Afican variant. Where were they collected? Does it make sense?

```
def get_sample_details(id):
    source_features_xpath = \
        "GBSeq/GBSeq_feature-table/GBFeature/GBFeature_key[text() = 'source']/parent::GBFeature/GBFeature_quals"
    country_xpath = \
        f"{source_features_xpath}/GBQualifier/GBQualifier_name[text() = 'country']/parent::GBQualifier/GBQualifier_value/text()"
    collection_date_xpath = \
        f"{source_features_xpath}/GBQualifier/GBQualifier_name[text() = 'collection_date']/parent::GBQualifier/GBQualifier_value

    handle = Entrez.efetch(db="nucleotide", id=id, rettype="gb", retmode="xml", email='no@one.com')
    seq_xml_string = handle.read()
    handle.close()

    root = ET.fromstring(seq_xml_string)

    found_nodes = root.xpath(country_xpath)
    if len(found_nodes) > 1:
        country_or_code = ValueError('More than one matching node found in XML for "country".')
    elif len(found_nodes) == 0:
        country_or_code = ValueError('"country" node not found in XML.')
    else:
        country_or_code = found_nodes[0].split(':')[0]

    try:
        country = countries.get(country_or_code).name
    except:
        country = country_or_code

    found_nodes = root.xpath(collection_date_xpath)
    if len(found_nodes) > 1:
        collection_date_string = ValueError('More than one matching node found in XML for "collection_date".')
```

```
    elif len(found_nodes) == 0:
        collection_date_string = ValueError('"collection_date" node not found in XML.')
    else:
        collection_date_string = found_nodes[0]

    try:
        collection_date = datetime.fromisoformat(collection_date_string)
    except:
        collection_date = collection_date_string

    return country, collection_date
```

```
sample_country_of_origin = defaultdict(list)
sample_collection_month = defaultdict(list)
unclassified_country, unclassified_date = [], []

# We can use the data in 'complete_covid_sa_sequences' (search area = 101, search sequence length = 11),
# as the last result there is a superset of all other results.
for seq_id in complete_covid_sa_sequences:
    try:
        country, collection_date = get_sample_details(seq_id)

        if isinstance(country, ValueError):
            country = None
        if isinstance(collection_date, ValueError):
            collection_date = None

    except Exception as e:
        print(e)
        country, collection_date = None, None

    (sample_country_of_origin[country] if country is not None else unclassified_country).append(seq_id)
    (sample_collection_month[str(collection_date)[:7]] if collection_date is not None else unclassified_date).append(seq_id)
```

```
unclassified_country
```

```
[]
```

```
unclassified_date
```

```
['HG999935.1', 'HG999939.1']
```

```
print('SA variant samples by country:\n-----------------------------------------------------\n')
sorted_samples_by_country = sorted(sample_country_of_origin.items(), key=lambda x: len(x[1]), reverse=True)
for country, samples in sorted_samples_by_country:
    print(country, f"({len(samples)})")
    for start_idx in range(0, len(samples), 5):
        print(", ".join(samples[start_idx:start_idx + 5]))
    print()
```

```
SA variant samples by country:
-----------------------------------------------------
```

```
United States of America (56)
MW912490.1, MW913437.1, MW914008.1, MW914015.1, MW914445.1
MW914542.1, MW905790.1, MW905844.1, MW905875.1, MW906028.1
MW906061.1, MW907199.1, MW907323.1, MW908243.1, MW908815.1
MW909222.1, MW909347.1, MW898265.1, MW883160.1, MW877177.1
MW880890.1, MW869153.1, MW844243.1, MW844258.1, MW844743.1
MW847393.1, MW849825.1, MW795351.1, MW842292.1, MW807936.1
MW808030.1, MW808712.1, MW809047.1, MW803575.1, MW803586.1
MW793569.1, MW796654.1, MW796843.1, MW792684.1, MW792756.1
MW793005.1, MW773815.1, MW763124.1, MW763125.1, MW763126.1
MW763127.1, MW763128.1, MW763129.1, MW763130.1, MW763131.1
MW763132.1, MW687146.1, MW617734.1, MW621453.1, MW580574.1
MW580576.1

Ghana (3)
MW598408.1, MW598413.1, MW598419.1

United Kingdom (2)
HG999935.1, HG999939.1

Germany (1)
MW822592.1

Italy (1)
MW789246.1

France (1)
MW580244.1
```

```
print('SA variant samples by collection month:\n-----------------------------------------------------\n')
sorted_samples_by_date = sorted(sample_collection_month.items(), key=lambda x: len(x[1]), reverse=True)
for country, samples in sorted_samples_by_date:
    print(country, f"({len(samples)})")
    for start_idx in range(0, len(samples), 5):
        print(", ".join(samples[start_idx:start_idx + 5]))
    print()
```

```
SA variant samples by collection month:
-----------------------------------------------------

2021-03 (29)
MW912490.1, MW913437.1, MW914008.1, MW914015.1, MW914445.1
MW914542.1, MW905790.1, MW905844.1, MW905875.1, MW906028.1
MW906061.1, MW907199.1, MW907323.1, MW908243.1, MW908815.1
MW909222.1, MW909347.1, MW898265.1, MW883160.1, MW877177.1
MW880890.1, MW869153.1, MW847393.1, MW849825.1, MW795351.1
MW808712.1, MW809047.1, MW796843.1, MW789246.1

2021-02 (26)
MW844243.1, MW844258.1, MW844743.1, MW842292.1, MW807936.1
MW808030.1, MW803575.1, MW803586.1, MW793569.1, MW796654.1
MW792684.1, MW792756.1, MW793005.1, MW773815.1, MW763124.1
MW763125.1, MW763126.1, MW763127.1, MW763128.1, MW763129.1
MW763130.1, MW763131.1, MW763132.1, MW687146.1, MW621453.1
MW580576.1

2021-01 (7)
MW822592.1, MW580244.1, MW617734.1, MW598408.1, MW598413.1
```

```
MW598419.1, MW580574.1
```

There are no samples that could not be classified by country, but for 2 samples from the UK we have no collection date.

We will not ignore these 2 samples, just to be on the safe side (as we have no info supporting we could ignore them).

The results make sense because:

1. All of the samples were found during 2021, after the SA variant has been identified to exist (with the exception of the 2 samples from the UK, for which we don't know the collection date).
2. None of the SA variant samples were found in China.
3. Some were found in Ghana.
4. Many were found in the US where the variant has been spotted according to news. Many were spotted there (presumably) due to more availability of sequencing technologies near by.

We would actually expect to find more in Ghana (or other African countries), however we presume this is not the case due to the lack of availability of sequencing technologies in that region of the world, combined with the difficulty of transporting samples to the US for sequencing.

In the following section we will actually cull some samples based on their protein structure.

17

Develop a method to extract the Spike sequence of one of the entries, translate, and report the protein sequence (hint: the South African spike is not very different than the Wuhan strain and must start with AUG)

We will use the start of the spike protein from the Wuhan variant to find the beginning of the SA variant's spike protein.

```
# We use the Wuhan variant's spike protein start that we have from a previous question.
```

```
spike_protein_nt_start = 'AUGUUUGUUUUUCUUGUUUUAUUGCCACUAGUCUCUAGUCAGUGUGUUAAUCUUACAACCAGAACUCAAUUACCCCCUGCAUACACUAAUUCUUUCACACGU
```

```
for sample_id in complete_covid_sa_sequences:
    finds = re.findall(spike_protein_nt_start, str(complete_covid_sa_sequences[sample_id].seq.transcribe()))
    print(sample_id, len(finds))
```

```
MW912490.1 1
MW913437.1 1
MW914008.1 1
MW914015.1 1
MW914445.1 1
MW914542.1 1
MW905790.1 1
MW905844.1 0
MW905875.1 0
MW906028.1 0
MW906061.1 0
MW907199.1 1
MW907323.1 1
MW908243.1 1
MW908815.1 1
MW909222.1 1
MW909347.1 1
MW898265.1 1
MW883160.1 1
MW877177.1 0
MW880890.1 1
MW869153.1 1
MW844243.1 1
MW844258.1 1
MW844743.1 1
MW847393.1 1
MW849825.1 1
MW795351.1 0
MW842292.1 1
MW822592.1 1
HG999935.1 1
HG999939.1 1
MW807936.1 0
MW808030.1 0
MW808712.1 0
MW809047.1 1
MW803575.1 1
MW803586.1 1
MW793569.1 1
MW796654.1 1
MW796843.1 1
MW789246.1 0
MW792684.1 1
MW792756.1 0
MW793005.1 0
MW773815.1 1
MW763124.1 1
MW763125.1 1
MW763126.1 1
MW763127.1 1
MW763128.1 1
```

```
MW763129.1 1
MW763130.1 1
MW763131.1 1
MW763132.1 1
MW687146.1 1
MW580244.1 0
MW617734.1 0
MW621453.1 1
MW598408.1 0
MW598413.1 0
MW598419.1 1
MW580574.1 0
MW580576.1 0
```

Trying to use the long sequence above doesn't yield a result for every sample (meaning we didn't find the spike start in every sample that we marked as belonging to the SA variant, probably due to more mutations/sequencing errors).

We'll try a shorter match.

```
spike_protein_nt_start = 'AUGUUUGUUUUUCUUGUUUUAUUGCCACUAGUCU'
```

```
for sample_id in complete_covid_sa_sequences:
    finds = re.findall(spike_protein_nt_start, str(complete_covid_sa_sequences[sample_id].seq.transcribe()))
    print(sample_id, len(finds))
```

```
MW912490.1 1
MW913437.1 1
MW914008.1 1
MW914015.1 1
MW914445.1 1
MW914542.1 1
MW905790.1 1
MW905844.1 1
MW905875.1 1
MW906028.1 1
MW906061.1 1
MW907199.1 1
MW907323.1 1
MW908243.1 1
MW908815.1 1
MW909222.1 1
MW909347.1 1
MW898265.1 1
MW883160.1 1
MW877177.1 1
MW880890.1 1
MW869153.1 1
MW844243.1 1
MW844258.1 1
MW844743.1 1
MW847393.1 1
```

```
MW849825.1 1
MW795351.1 1
MW842292.1 1
MW822592.1 1
HG999935.1 1
HG999939.1 1
MW807936.1 1
MW808030.1 1
MW808712.1 1
MW809047.1 1
MW803575.1 1
MW803586.1 1
MW793569.1 1
MW796654.1 1
MW796843.1 1
MW789246.1 1
MW792684.1 1
MW792756.1 1
MW793005.1 1
MW773815.1 1
MW763124.1 1
MW763125.1 1
MW763126.1 1
MW763127.1 1
MW763128.1 1
MW763129.1 1
MW763130.1 1
MW763131.1 1
MW763132.1 1
MW687146.1 1
MW580244.1 1
MW617734.1 1
MW621453.1 1
MW598408.1 1
MW598413.1 1
MW598419.1 1
MW580574.1 1
MW580576.1 1
```

This yields exactly one match per sample, so we can use this to identify the protein start.

We will use this to get the spike protein for each of the samples we marked as the SA variant.

```
# Map between sample id and spike protein sequence.
sa_spike_proteins = {}
for sample_id in complete_covid_sa_sequences:
    sample_spike_start = re.search(spike_protein_nt_start, str(complete_covid_sa_sequences[sample_id].seq.transcribe())).span()
    sa_spike_proteins[sample_id] = get_translated_region(complete_covid_sa_sequences[sample_id].seq.transcribe()[sample_spike_st
```

Let's have a look at all the end codons of all the samples.

Perhaps we can find a distinction between the SA and Wuhan variants based on that?

```
set([d['End Codon'] for d in sa_spike_proteins.values()])
```

```
{'UAA'}
```

As all the samples have the same end codon, we can't distinguish between the variants based on them.

```
# Reverse map between the spike protein sequence and the sample ids that correspond to it.
unique_sa_spike_proteins = defaultdict(list)
for sample_id in sa_spike_proteins:
    unique_sa_spike_proteins[str(sa_spike_proteins[sample_id]['Amino Seq'])].append(sample_id)


# Histogram proteins explain how many samples each.
number_of_samples_per_unique_sa_spike_protein_hist = defaultdict(int)

# Unique spike proteins that don't contain coding ambiguity (that is marked by a non existing 'X' amino).
unique_sa_spike_proteins_without_coding_ambiguity = []
culled_sa_spike_proteins = []

for unique_sa_spike_protein in unique_sa_spike_proteins:
    number_of_samples_per_unique_sa_spike_protein_hist[len(unique_sa_spike_proteins[unique_sa_spike_protein])] += 1

    # According to: https://www.genome.jp/kegg/catalog/codes1.html
    if 'X' not in unique_sa_spike_protein and 'B' not in unique_sa_spike_protein \
        and 'Z' not in unique_sa_spike_protein and 'J' not in unique_sa_spike_protein:
        unique_sa_spike_proteins_without_coding_ambiguity.append(unique_sa_spike_protein)
    else:
        culled_sa_spike_proteins.append(unique_sa_spike_protein)


number_of_samples_per_unique_sa_spike_protein_hist
```

```
defaultdict(int, {1: 26, 2: 3, 3: 2, 8: 1, 18: 1})
```

```
len(unique_sa_spike_proteins), len(unique_sa_spike_proteins_without_coding_ambiguity)
```

```
(33, 10)
```

We found that our 64 samples, are described by a total of 33 different spike proteins, and more specifically:

- 18 of the samples are described by one protein.

- 8 of the samples are described by a different protein.
- 6 of the samples are described by 2 other different proteins (each describes 3 samples).
- 6 of the samples are described by 3 other different proteins (each protein describing a pair of samples).
- And 26 samples are described by a unique protein.

Then, we cull samples that have amino coding ambiguities (that have any of the non-existing **X, B, J, Z** aminos in the translated amino sequence), and when we look at the number of unique spike proteins left to analyze, we remain with 10.

We could have still used these samples while ignoring only the non-existing aminos instead of the entrire sample, however in the following section we will be using majority-vote (to decide which amino the SA variant virus' spike protein has in cases of disagreement between the samples), and using part of the sequence would mean the majority vote would be based on a different number of "voters" per amino acid.

We will however look at the proteins we culled as well (in a dedicated sub-section at the end of this section), and see if taking them into account produces different results.

The following **35** samples are used (divided by their spike proteins):

```
sorted_samples = sorted(
    {protein: samples for protein, samples in unique_sa_spike_proteins.items() if protein in unique_sa_spike_proteins_without_c
    key=lambda x: len(x[1]), reverse=True)

for country, samples in sorted_samples:
    for start_idx in range(0, len(samples), 5):
        print(", ".join(samples[start_idx:start_idx + 5]))
    print()
```

```
MW913437.1, MW914008.1, MW914015.1, MW914542.1, MW905790.1
MW907199.1, MW907323.1, MW908243.1, MW909222.1, MW909347.1
MW844243.1, MW844743.1, MW847393.1, MW809047.1, MW796654.1
MW792684.1, MW773815.1, MW598419.1

MW905844.1, MW877177.1, MW795351.1, MW789246.1, MW793005.1
MW580244.1, MW598408.1, MW598413.1

MW808712.1, MW617734.1

MW914445.1
```

```
MW822592.1

MW808030.1

MW796843.1

MW792756.1

MW763124.1

MW621453.1
```

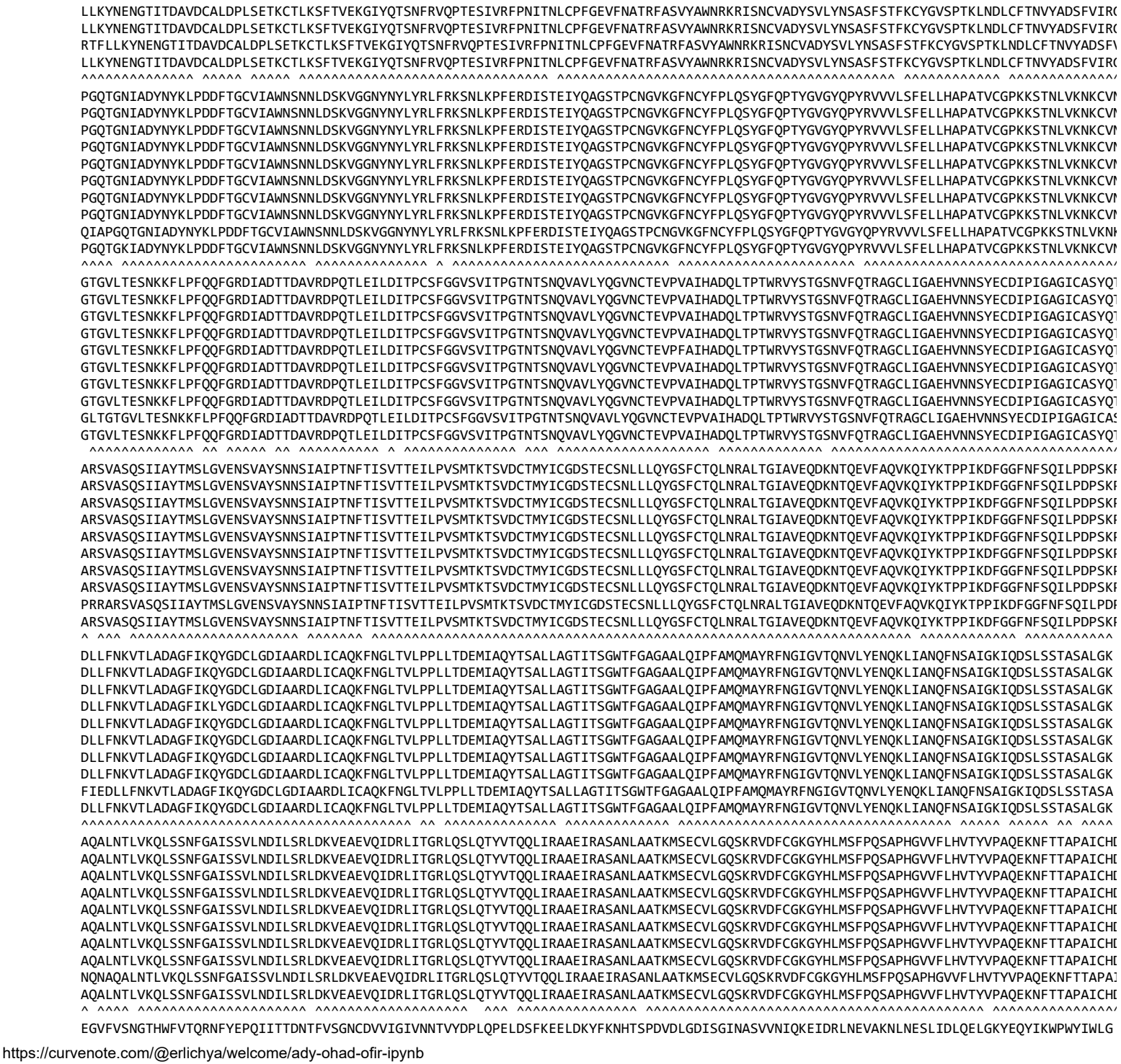
Let's look at how different the 10 amino sequences are:

```python
s1 = unique_sa_spike_proteins_without_coding_ambiguity[0]

step = 136
for start_idx in range(0, len(s1), step):
    for s in unique_sa_spike_proteins_without_coding_ambiguity:
        print(s[start_idx:start_idx + step])

    s_parts = [s[start_idx:start_idx + step] for s in unique_sa_spike_proteins_without_coding_ambiguity]
    s_is = [set([s_part[i] for s_part in s_parts]) for i in range(len(s_parts[0]))]
    comparison = [' ' if len(s_i) == 1 else '^' for s_i in s_is]
    print("".join(comparison))
```

```
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLITRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
                 ^^
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITKFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWSAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
                                                                                                    ^    ^^^^^^^^^^^^^^^^^^ ^^^^
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
```



```
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
^^^^^^^^^^^^^^ ^^^^^ ^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^ ^^^^^^^^^^^^^

PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
PGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
^^^^ ^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^ ^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPFAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
 ^^^^^^^^^^^^^ ^^ ^^^^^ ^^ ^^^^^^^^^^ ^ ^^^^^^^^^^^^^^ ^^^ ^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSK
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSK
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSK
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSK
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSK
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSK
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSK
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSK
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSK
^ ^^^ ^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^ ^^^^^^^^^^^

DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKLYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^ ^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^ ^^^^^ ^^ ^^^^

AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
^ ^^^^ ^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^  ^^^ ^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^

EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
```

```
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQEFGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIW
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^ ^^^^^^^^^^ ^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^ ^^^^^^^ ^^^
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
^^^^^^^^^^^^^^  ^^^^^^ ^  ^^^^^^ ^^^^^^^^ ^^^^
```

We see that while most proteins agree in most cases on the amino level, one protein is "vastly" different. Actually, it contains a **LAL** amino sequence that the other samples don't, so we will try to "move" the other sequences by 3 aminos at that point to see how similar the proteins are then.

```python
different_amino_seq = 'LALHRSYLTPGDSSSGW'
regular_amino_seq = 'HRSYLTPGDSSSGW'

unique_moved_sa_spike_proteins_without_coding_ambiguity = []

for s in unique_sa_spike_proteins_without_coding_ambiguity:
    if s.find(different_amino_seq) != -1:
        unique_moved_sa_spike_proteins_without_coding_ambiguity.append(s)
        weird_sa_spike_proteins_without_coding_ambiguity = s
    else:
        difference_start = s.find(regular_amino_seq)
        unique_moved_sa_spike_proteins_without_coding_ambiguity.append(f'{s[:difference_start]}   {s[difference_start:]}')

s1 = unique_sa_spike_proteins_without_coding_ambiguity[0]

step = 136
for start_idx in range(0, len(s1), step):
    for s in unique_moved_sa_spike_proteins_without_coding_ambiguity:
        print(s[start_idx:start_idx + step])

    s_parts = [s[start_idx:start_idx + step] for s in unique_moved_sa_spike_proteins_without_coding_ambiguity]
    s_is = [set([s_part[i] for s_part in s_parts]) for i in range(len(s_parts[0]))]
    comparison = [' ' if len(s_i) == 1 else '^' for s_i in s_is]
    print("".join(comparison))
```

```
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLITRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
                 ^^
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTL   HRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITKFQTL   HRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTL   HRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTL   HRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTL   HRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTL   HRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTL   HRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTL   HRSYLTPGDSSSGWSAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTL   HRSYLTPGDSSSGWTAGAAA
                                                                                                    ^    ^^^              ^
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF

QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
        ^
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPFAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
                                                                             ^
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGNFSQILPD
```

```
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD

FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKLYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
                   ^
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA

FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQEFGKYEQYIKWPWYI
FPLEGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
  ^                                                                                                               ^
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

```
wuhan_sars_cov_2_data['Amino Seq'].find(different_amino_seq)
```

```
241
```

```
unique_sa_spike_proteins[weird_sa_spike_proteins_without_coding_ambiguity]
```

```
['MW763124.1']
```

We see that now the proteins are much more similar.

On closer inspection, we find that:

1. There is only one sample that has that specific protein.
2. The **LAL** amino sequence (together with the following 14 aminos) does not appear in any of the other SA variant samples, but it does appear in the Wuhan variant, and in its vaccine.

We conclude that that protein and corresponding sample belong to the Wuhan variant, and will not be using it for the SA vaccine.

```
unique_sa_spike_proteins_without_coding_ambiguity.remove(weird_sa_spike_proteins_without_coding_ambiguity)
```

After removing the problematic sample we are left with the following **34** sequences (again divided by their spike protein):

```
sorted_samples = sorted(
    {protein: samples for protein, samples in unique_sa_spike_proteins.items() if protein in unique_sa_spike_proteins_without_c
    key=lambda x: len(x[1]), reverse=True)

for country, samples in sorted_samples:
    for start_idx in range(0, len(samples), 5):
        print(", ".join(samples[start_idx:start_idx + 5]))
    print()
```

```
MW913437.1, MW914008.1, MW914015.1, MW914542.1, MW905790.1
MW907199.1, MW907323.1, MW908243.1, MW909222.1, MW909347.1
MW844243.1, MW844743.1, MW847393.1, MW809047.1, MW796654.1
MW792684.1, MW773815.1, MW598419.1

MW905844.1, MW877177.1, MW795351.1, MW789246.1, MW793005.1
MW580244.1, MW598408.1, MW598413.1

MW808712.1, MW617734.1

MW914445.1

MW822592.1

MW808030.1
```

```
MW796843.1

MW792756.1

MW621453.1
```

For the proteins we are using, we will use majority vote to decide what amino acid should be used for the vaccine.

This is done for 2 reasons:

1. If there's a sequencing error, we'll overcome it by using this method.
2. If it is not a sequencing error, it makes more sense to use the most common one for the vaccine.

```
print("On the left are the number of samples for that specific protein, at the bottom which amino to use in case of ambiguity")
print("-----------------------------------------------------------------------------------------------------------------------")

s1 = unique_sa_spike_proteins_without_coding_ambiguity[0]

maj_votes = []
for i in range(len(s1)):
    amino_maj_votes = defaultdict(int)
    for s in unique_sa_spike_proteins_without_coding_ambiguity:
        amino_maj_votes[s[i]] += len(unique_sa_spike_proteins[s])
    maj_votes.append(amino_maj_votes)

step = 136
for start_idx in range(0, len(s1), step):
    for s in unique_sa_spike_proteins_without_coding_ambiguity:
        print(f'{len(unique_sa_spike_proteins[s]):2}', s[start_idx:start_idx + step])

    s_parts = [s[start_idx:start_idx + step] for s in unique_sa_spike_proteins_without_coding_ambiguity]
    s_is = [[s_part[i] for s_part in s_parts] for i in range(len(s_parts[0]))]
    comparison = [' ' if len(maj_vote) == 1 else max(maj_vote, key=maj_vote.get) for maj_vote in maj_votes[start_idx:start_idx +
    print("  ", "".join(comparison), '\n')
```

```
On the left are the number of samples for that specific protein, at the bottom which amino to use in case of ambiguity
-----------------------------------------------------------------------------------------------------------------------

18 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN\
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN\
 8 MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN\
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN\
 1 MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN\
 2 MFVFLVLLPLVSSQCVNLITRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN\
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN\
 1 MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN\
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN\
                   LT
```

```
18 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAY
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITKFQTLHRSYLTPGDSSSGWTAGAAAY
 8 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAY
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAY
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAY
 2 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAY
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAY
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWSAGAAAY
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAY
                                                                                                       R                  T

18 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
 1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
 8 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
 1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
 1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
 2 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
 1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
 1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
 1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV


18 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
 1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
 8 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
 1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
 1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
 2 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
 1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
 1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
 1 PGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
        N

18 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICAS
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICAS
 8 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICAS
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICAS
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPFAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICAS
 2 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICAS
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICAS
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICAS
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICAS
                                                                             V

18 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP
 8 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP
 2 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP


18 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 8 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKLYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
```

```
 2 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
                    Q

18 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 8 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 2 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA


18 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 8 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 2 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQEFGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
                                                                                                                            L

18 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 8 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 2 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

So the spike protein we will use as the base for our vaccine is:

```python
sa_vaccine_base = "".join([max(maj_vote, key=maj_vote.get) for maj_vote in maj_votes])
```

```python
step = 136

for start_idx in range(0, len(sa_vaccine_base), step):
    print(sa_vaccine_base[start_idx:start_idx + step])
```

```
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
```

```
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

We also find the majority vote on the nucleotide level (among the proteins that agreed with the majority vote on the amino level).

```
# # Sample to data (including nucleotides)
# sa_spike_proteins.keys()

# # Protein to sample
# unique_sa_spike_proteins.keys()

# # Proteins analyzed
# unique_sa_spike_proteins_without_coding_ambiguity

sa_vaccine_base_nts = ""

for amino_idx, amino in enumerate(sa_vaccine_base):
    codons_for_amino = defaultdict(int)

    for protein in unique_sa_spike_proteins_without_coding_ambiguity:
        if protein[amino_idx] == amino:
            for sample_id_with_protein in unique_sa_spike_proteins[protein]:
                codon_for_amino_in_sample = sa_spike_proteins[sample_id_with_protein]['Nucleotide Seq'][amino_idx * 3:(amino_id
                codons_for_amino[codon_for_amino_in_sample] += 1

    selected_codon = max(codons_for_amino, key=codons_for_amino.get)
    sa_vaccine_base_nts += selected_codon
```

Yaniv Erlich
11:20 PM, May 12

Very nice work and correct!

Quick detour:

How do the proteins that were discarded because of sequencing ambiguities look, and how do these affect the result?

```
s1 = culled_sa_spike_proteins[0]

step = 136
for start_idx in range(0, len(s1), step):
    for s in culled_sa_spike_proteins:
        print(s[start_idx:start_idx + step])
```

```
    s_parts = [s[start_idx:start_idx + step] for s in culled_sa_spike_proteins]
    s_is = [set([s_part[i] for s_part in s_parts]) for i in range(len(s_parts[0]))]
    comparison = [' ' if len(s_i) == 1 else '^' for s_i in s_is]
    print("".join(comparison))
```

```
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPSYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYXXXKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLITRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
                 ^^       ^          ^^^
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAXXX
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSXXXXXXXXXXXXX
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSXXXXXXXXXXXXX
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTXXXAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTXXXXXYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGXXXXXX
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGXXXXXX
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSXXXXXXXXXXXXX
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTXGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTXXXLHRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFXXXXXXHRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFXXXXXLHRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTXXXXXXSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTXXXXXRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAXXXXXXX
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYV
                                                                                                      ^^^^^^^^^^^^^^^^^^^^^^^^^^
LLKYNXNGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
XXXXNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
XXXXNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCXXXSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITXAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
XXXXXXXXXXXXXXXXXXDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
```

```
XXXXXXXXXXXXAVDXXXXXXXXXXXXXXXXXVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
XXXXXXXXXXXXXXDXXXXXXXXXXXXXXXXXXXXGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
XXXXNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYXXXXXXXXXXDXXXXXXXXXXXXXXXXXXXXXXYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTXXXAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
XXXXXXXXXXXXAVXXXXXXXXXXKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALXXXXXXXXXXXXXXXXXXXXXXXXXXXXPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^ ^^^^^^^^^^^^^

PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGXIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
^^^^ ^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^ ^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
```

```
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQT
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQT
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQT
 ^^^^^^^^^^^^^ ^^ ^^^^^ ^^ ^^^^^^^^^^ ^ ^^^^^^^^^^^^^^ ^^^ ^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
^ ^^^ ^^^^^^^^^^^^^^^^^^^^ ^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^ ^^^^^^^^^^^

DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFXAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYNSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGXXXXXXXXXDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLXXXXTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^ ^^^^^ ^^ ^^^^

AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
```

```
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
^ ^^^^ ^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^  ^^^ ^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^

EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNDSLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDXXXXXXXXXXXXXXXXQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^ ^^^^^ ^^^^^^^ ^^

VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

```
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
^^^^^^^^^^^^^^  ^^^^^^ ^  ^^^^^^ ^^^^^^^^ ^^^^
```

We see that at least some of the proteins discarded, most probably belong to the Wuhan variant (are longer by 3 amino acids, that can probably be attributed to the extra **LAL** amino chain that the Wuhan variant has when compared to the SA variant).

Let's see what result we'd have got, had we not dropped all the sequences with sequencing ambiguities, and instead just dropped only those that are longer by 3 aminos and probably belong to the SA variant:

```
wuhan_length = len(sa_spike_proteins[unique_sa_spike_proteins[weird_sa_spike_proteins_without_coding_ambiguity][0]]['Amino Seq'
```

```
# Histogram proteins explain how many samples each.
number_of_samples_per_unique_sa_spike_protein_hist = defaultdict(int)

# Unique spike proteins that don't contain coding ambiguity (that is marked by a non existing 'X' amino).
unique_sa_spike_proteins_without_coding_ambiguity = []
culled_sa_spike_proteins = []

for unique_sa_spike_protein in unique_sa_spike_proteins:
    number_of_samples_per_unique_sa_spike_protein_hist[len(unique_sa_spike_proteins[unique_sa_spike_protein])] += 1

    # According to: https://www.genome.jp/kegg/catalog/codes1.html

    if len(unique_sa_spike_protein) != wuhan_length:
        unique_sa_spike_proteins_without_coding_ambiguity.append(unique_sa_spike_protein)
    else:
        culled_sa_spike_proteins.append(unique_sa_spike_protein)
```

```
number_of_samples_per_unique_sa_spike_protein_hist
```

```
defaultdict(int, {1: 26, 2: 3, 3: 2, 8: 1, 18: 1})
```

```
len(unique_sa_spike_proteins), len(unique_sa_spike_proteins_without_coding_ambiguity)
```

```
(33, 27)
```

The numbers we found here are different, but that makes sense as every sample which has sequencing ambiguities will have a different protein than all the rest of the samples (unless the ambiguities are in exactly the same places, in which case several samples might be explained by a single protein).

In this case, the following samples **55** samples are used (divided by their spike proteins):

```
sorted_samples = sorted(
    {protein: samples for protein, samples in unique_sa_spike_proteins.items() if protein in unique_sa_spike_proteins_without_c
    key=lambda x: len(x[1]), reverse=True)

for country, samples in sorted_samples:
    for start_idx in range(0, len(samples), 5):
        print(", ".join(samples[start_idx:start_idx + 5]))
    print()
```

```
MW913437.1, MW914008.1, MW914015.1, MW914542.1, MW905790.1
MW907199.1, MW907323.1, MW908243.1, MW909222.1, MW909347.1
MW844243.1, MW844743.1, MW847393.1, MW809047.1, MW796654.1
MW792684.1, MW773815.1, MW598419.1

MW905844.1, MW877177.1, MW795351.1, MW789246.1, MW793005.1
MW580244.1, MW598408.1, MW598413.1

MW842292.1, MW803575.1, MW803586.1

MW906028.1, MW906061.1

MW808712.1, MW617734.1

MW912490.1

MW914445.1

MW905875.1

MW908815.1

MW898265.1

MW883160.1

MW880890.1

MW869153.1

MW844258.1

MW849825.1
```

```
MW822592.1

HG999935.1

HG999939.1

MW807936.1

MW808030.1

MW793569.1

MW796843.1

MW792756.1

MW687146.1

MW621453.1

MW580574.1

MW580576.1
```

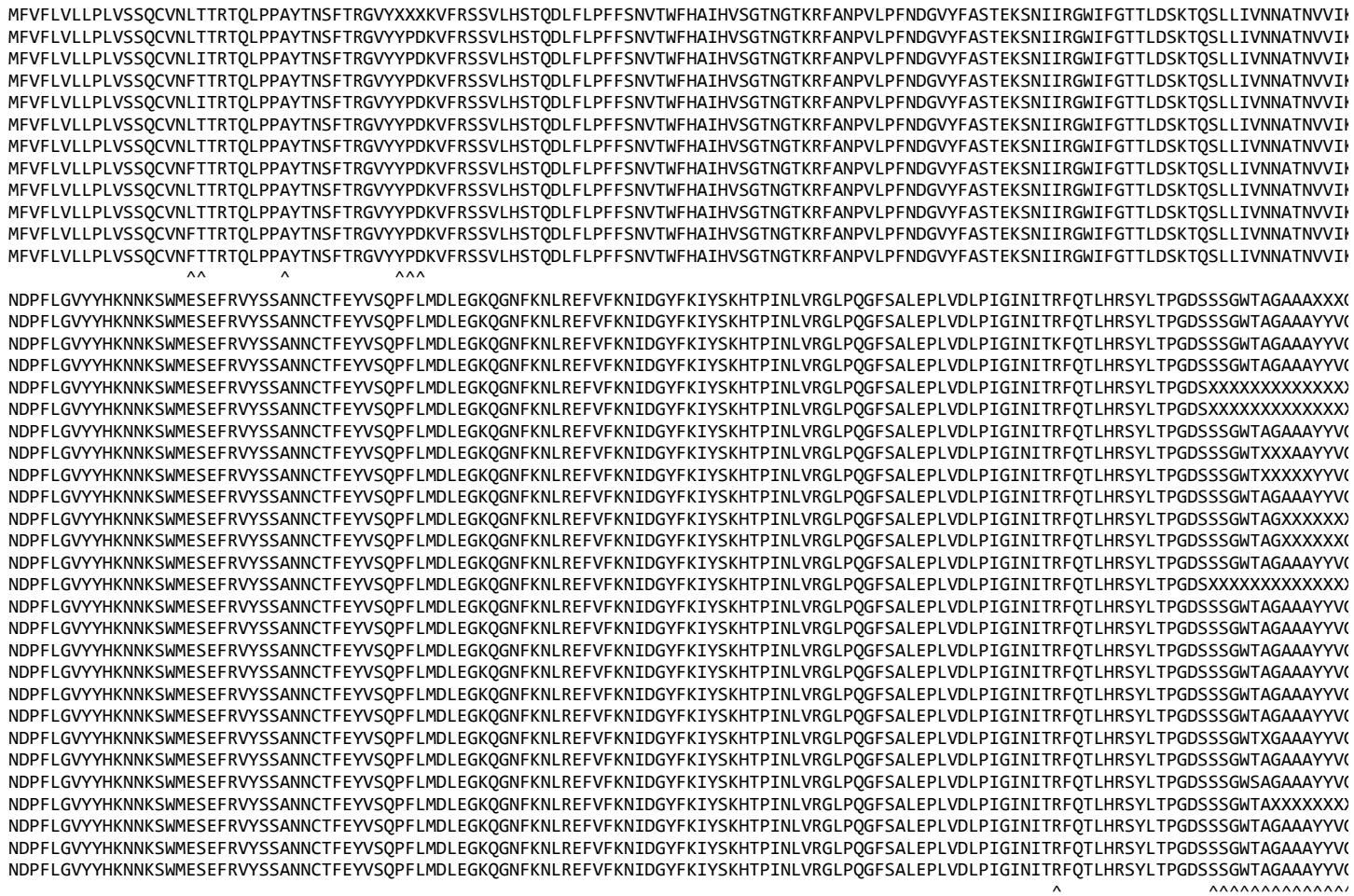
Let's look at how different the amino sequences are now:

```python
s1 = unique_sa_spike_proteins_without_coding_ambiguity[0]

step = 136
for start_idx in range(0, len(s1), step):
    for s in unique_sa_spike_proteins_without_coding_ambiguity:
        print(s[start_idx:start_idx + step])

    s_parts = [s[start_idx:start_idx + step] for s in unique_sa_spike_proteins_without_coding_ambiguity]
    s_is = [set([s_part[i] for s_part in s_parts]) for i in range(len(s_parts[0]))]
    comparison = [' ' if len(s_i) == 1 else '^' for s_i in s_is]
    print("".join(comparison))
```

```
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPSYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
```



```
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYXXXKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIK
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIK
MFVFLVLLPLVSSQCVNLITRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIK
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIK
MFVFLVLLPLVSSQCVNLITRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIK
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIK
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIK
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIK
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIK
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIK
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIK
MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIK
                 ^^       ^          ^^^
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAXXXG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITKFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSXXXXXXXXXXXXXX
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSXXXXXXXXXXXXXX
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTXXXAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTXXXXXYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGXXXXXXX
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGXXXXXXG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSXXXXXXXXXXXXXX
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTXGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWSAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAXXXXXXXX
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVG
                                                                                                    ^              ^^^^^^^^^^^^^^
LLKYNXNGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
XXXXNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
XXXXNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
LLKYNENGTITDAVDCALDPLSETKCXXXSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
LLKYNENGTITXAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
XXXXXXXXXXXXXXXXXXDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
XXXXXXXXXXXXAVDXXXXXXXXXXXXXXXXXVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
XXXXXXXXXXXXXXDXXXXXXXXXXXXXXXXXXXXGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
XXXXNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
LLKYXXXXXXXXXXDXXXXXXXXXXXXXXXXXXXXXXYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRG
```

```
LLKYNENGTXXXAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
XXXXXXXXXXXXAVXXXXXXXXXXKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
LLKYNENGTITDAVDCALXXXXXXXXXXXXXXXXXXXXXXXXXXXXPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGXIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCV
     ^

GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPFAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQ
```

```
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQT
GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQT
                                                                          ^
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP
ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKP

DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFXAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYNSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKLYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGXXXXXXXXXDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLXXXXTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGK
                ^                     ^^^^^^^^^       ^              ^
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
```

```
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD
AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHD

EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNDSLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDXXXXXXXXXXXXXXXXQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQEFGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLG
                                                                        ^^^^^^^^^^^^^^^^               ^       ^

VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

```
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

All the proteins look very similar, except (mostly) for the places with sequencing ambiguities.

For the proteins we are using, we will again use majority vote to decide what amino acid should be used for the vaccine, however ignore the "voters" that are ambiguous (per amino acid).

```python
print("On the left are the number of samples for that specific protein, at the bottom which amino to use in case of ambiguity")
print("--------------------------------------------------------------------------------------------------------------------------

s1 = unique_sa_spike_proteins_without_coding_ambiguity[0]

maj_votes = []
for i in range(len(s1)):
    amino_maj_votes = defaultdict(int)
    for s in unique_sa_spike_proteins_without_coding_ambiguity:
        if s[i] not in 'XBZJ':
            amino_maj_votes[s[i]] += len(unique_sa_spike_proteins[s])
    maj_votes.append(amino_maj_votes)

step = 136
for start_idx in range(0, len(s1), step):
    for s in unique_sa_spike_proteins_without_coding_ambiguity:
        print(f'{len(unique_sa_spike_proteins[s]):2}', s[start_idx:start_idx + step])

    s_parts = [s[start_idx:start_idx + step] for s in unique_sa_spike_proteins_without_coding_ambiguity]
    s_is = [[s_part[i] for s_part in s_parts] for i in range(len(s_parts[0]))]
    comparison = [' ' if len(maj_vote) == 1 else max(maj_vote, key=maj_vote.get) for maj_vote in maj_votes[start_idx:start_idx
    print("   ", "".join(comparison), '\n')
```

```
On the left are the number of samples for that specific protein, at the bottom which amino to use in case of ambiguity
-------------------------------------------------------------------------------------------------------------------------

 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
18 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 8 MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 2 MFVFLVLLPLVSSQCVNLTTRTQLPPSYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 3 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYXXXKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLITRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 2 MFVFLVLLPLVSSQCVNLITRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
 1 MFVFLVLLPLVSSQCVNFTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATN
                    LT       A

 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
18 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITKFQTLHRSYLTPGDSSSGWTAGAAA
 8 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSXXXXXXXXXX
 2 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSXXXXXXXXXX
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTXXXAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTXXXXX
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGXXX
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGXXX
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
 3 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSXXXXXXXXXX
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
 2 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTXGAAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWSAGAAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAXXXX
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
 1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAA
```

```
  1 NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAY
                                                                                                        R                 T

  1 LLKYNXNGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
 18 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  8 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 XXXXNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  2 XXXXNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITDAVDCALDPLSETKCXXXSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITXAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 XXXXXXXXXXXXXXXXXXDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 XXXXXXXXXXXXAVDXXXXXXXXXXXXXXXXXVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 XXXXXXXXXXXXXXDXXXXXXXXXXXXXXXXXXXXGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  3 XXXXNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYXXXXXXXXXXDXXXXXXXXXXXXXXXXXXXXXXYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  2 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTXXXAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 XXXXXXXXXXXXAVXXXXXXXXXXKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITDAVDXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV
  1 LLKYNENGTITDAVDCALXXXXXXXXXXXXXXXXXXXXXXXXXXXXPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFV


  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
 18 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  8 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  2 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGXIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  3 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  2 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
  1 PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNK
         N
```

```
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
18 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 8 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 2 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 3 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPFAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 2 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
 1 GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
                                                                             V

 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
18 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 8 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 2 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 3 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 2 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
 1 ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD


 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
18 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
```

```
 8 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 2 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFXAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYNSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 3 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKLYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGXXXXXXXXXDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLXXXXTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 2 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
 1 DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
                   Q                                     T

 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
18 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 8 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 2 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 3 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 2 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
 1 AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA


 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
18 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 8 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 2 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
```

```
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNDSLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 3 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDXXXXXXXXXXXXXXXXQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 2 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQEFGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
 1 EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
                                                                                                          E       L

 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
18 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 8 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 2 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 3 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 2 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
 1 VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

So the spike protein we will use as the base for our vaccine is:

```
sa_vaccine_base_with_ambiguous = "".join([max(maj_vote, key=maj_vote.get) for maj_vote in maj_votes])
```

```
sa_vaccine_base_with_ambiguous == sa_vaccine_base
```

```
True
```

Now do the same on the nucleotide level (again among the proteins that agreed with the majority vote on the amino level but ignoring sequencing ambiguities).

```
sa_vaccine_base_nts_with_ambiguous = ""

for amino_idx, amino in enumerate(sa_vaccine_base):
    codons_for_amino = defaultdict(int)

    for protein in unique_sa_spike_proteins_without_coding_ambiguity:
        if protein[amino_idx] == amino and protein[amino_idx] not in 'XBZJ':
            for sample_id_with_protein in unique_sa_spike_proteins[protein]:
                codon_for_amino_in_sample = sa_spike_proteins[sample_id_with_protein]['Nucleotide Seq'][amino_idx * 3:(amino_id
                codons_for_amino[codon_for_amino_in_sample] += 1

    selected_codon = max(codons_for_amino, key=codons_for_amino.get)
    sa_vaccine_base_nts_with_ambiguous += selected_codon
```

```
sa_vaccine_base_nts_with_ambiguous == sa_vaccine_base_nts
```

```
True
```

And so we found that even had we used the samples discarded because of sequencing ambiguities, but not using the specific places that were actually ambiguous, we would have got the same result, on both the amino and the nucleotide levels.

End of detour

18

Use everything you learned so far, including the look up table of Pfizer and the addition of special nucleotides to design a South African mRNA vaccine. Don't forget the add the 5'UTR and 3'UTR. Report the vaccine sequence.

```
vaccines_data[company_names[0]]
```

```
{'Amino Seq': Seq('MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDL...HYT'),
 'End Codon': 'UGA',
 'End Codon Index': 3873,
 'Nucleotide Seq': 'AUGUUCGUGUUCCUGGUGCUGCUGCCUCUGGUGUCCAGCCAGUGUGUGAACCUGACCACCAGAACACAGCUGCCUCCAGCCUACACCAACAGCUUUACCAGAGGCGUGU
 'Start Codon Index': 54}
```

```
len(vaccines_data[company_names[0]]['Amino Seq']), len(sa_vaccine_base),
```

```
(1273, 1270)
```

We see that the length of the Pfizer Wuhan vaccine's spike protein is different than that of the SA variant's.

Let's compare the spike proteins:

```
print('Vac: Pfizer vaccine  -  Vir: SA variant virus\n==========================================================================\n')

step = 68
for start_idx in range(0, len(vaccines_data[company_names[0]]['Amino Seq']), step):
    try:
        vac_part = vaccines_data[company_names[0]]['Amino Seq'][start_idx:start_idx + step]
        vir_part = sa_vaccine_base[start_idx:start_idx + step]

        print('Vac:', vac_part)
        print('Vir:', vir_part)
        print('    ', "".join([' ' if vac_part[i] == vir_part[i] else '^' for i in range(len(vac_part))]))
    # We ignore exceptions created by index out of bound.
    except IndexError:
        pass
```

```
Vac: Pfizer vaccine  -  Vir: SA variant virus
==========================================================================

Vac: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI
Vir: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI

Vac: HVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
Vir: HVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
                ^
Vac: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY
Vir: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY
```

```
Vac: SKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP
Vir: SKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTF
               ^                          ^^^^^^^^^^^^^^^^^^ ^^^^^ ^^^^^^

Vac: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE
Vir: LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFN
     ^^^^^^^^^^^^^ ^^^^^ ^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^

Vac: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR
Vir: ATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIA
     ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^

Vac: QIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
Vir: PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTP
     ^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

Vac: STPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
Vir: CNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLT
     ^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

Vac: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY
Vir: GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGV
      ^^^^^^^^^^^^^ ^^ ^^^^^ ^^ ^^^^^^^^^^ ^ ^^^^^^^^^^^^^^ ^^^ ^^^^^^^^^

Vac: QDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
Vir: NCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRR
     ^^^^^^^^^^ ^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

Vac: PRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
Vir: ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSN
     ^ ^^^ ^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

Vac: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS
Vir: LLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIE
     ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^^ ^^^^^^^^^^^   ^ ^^^

Vac: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS
Vir: DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWT
     ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^ ^^^^^^^^^^^^^^ ^^^^^^^^

Vac: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV
Vir: FGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQN
     ^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

Vac: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
Vir: AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASAN
     ^ ^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

Vac: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH
Vir: LAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPR
     ^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

Vac: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF
Vir: EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNH
     ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

Vac: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL
Vir: TSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAI
     ^^^^^^^^ ^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

Vac: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
Vir: VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

We observe the extra **LAL** aminos in the Pfizer vaccine that we discussed earlier.

We can align the sequences by jumping 3 aminos in the SA variant where the inconsistency begins:

```
# This we took from above
inconsistency_start = sa_vaccine_base.find('HRSYLTPGDSSSGWTAGAAAYYVGYLQPRTF')
```

```
modified_sa_vaccine_base_amino = f"{sa_vaccine_base[:inconsistency_start]}   {sa_vaccine_base[inconsistency_start:]}"
modified_sa_vaccine_base_codon = f"{sa_vaccine_base_nts[:inconsistency_start * 3]}         {sa_vaccine_base_nts[inconsistency_st

print('Vac: Pfizer vaccine  -  Vir: SA variant virus\n==========================================================================\n')

step = 68
for start_idx in range(0, len(vaccines_data[company_names[0]]['Amino Seq']), step):
    vac_part = vaccines_data[company_names[0]]['Amino Seq'][start_idx:start_idx + step]
    vir_part = modified_sa_vaccine_base_amino[start_idx:start_idx + step]

    print('Vac:', vac_part)
    print('Vir:', vir_part)
    print('    ', "".join([' ' if vac_part[i] == vir_part[i] else '^' for i in range(len(vac_part))]))
```

```
Vac: Pfizer vaccine  -  Vir: SA variant virus
==========================================================================

Vac: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI
Vir: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI

Vac: HVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
Vir: HVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
                ^
Vac: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY
Vir: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY

Vac: SKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP
Vir: SKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTL   HRSYLTPGDSSSGWTAGAAAYYVGYLQP
               ^                           ^^^
Vac: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE
Vir: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE

Vac: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR
Vir: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR

Vac: QIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
Vir: QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
            ^
Vac: STPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
Vir: STPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
           ^               ^
Vac: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY
Vir: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY

Vac: QDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
Vir: QGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
      ^
Vac: PRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
Vir: PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
                        ^
Vac: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS
Vir: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS

Vac: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS
Vir: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS

Vac: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV
Vir: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV
```

```
Vac: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
Vir: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
                                          ^^
Vac: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH
Vir: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH

Vac: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF
Vir: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF

Vac: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL
Vir: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL

Vac: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
Vir: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

Now that we have a spike protein comparable to that of the vaccine (and by that also to that of the Wuhan variant that is only different on 2 amino acids vs the vaccine), we will construct the vaccine for the SA variant using the following rules:

1. For every amino acid that we see both in the existing vaccine and in the SA virus variant, we will use the vaccine's codon.
2. For amino acids that are different between them, we will use a codon from the mapping table we created earlier, by the same distribution of target codons found there (trying to keep an overall GC content similar to that of the vaccine).
3. Amino acids that are "missing" in the SA variant will be ignored.
4. For the 2 amino acids that are different between the existing vaccine and the Wuhan variant, we will use the existing vaccine's codons.
5. We will use a UGA end codon (of the vaccine) instead of the UAA codon that the different virus variants have.
6. We will add the 5', and 3' UTRs of the vaccine.

The overarching logic behind these rules is that the Pfizer vaccine amino changes vs the Wuhan spike protein are intentional, and so take precedence (there are 2 such instances), but everywhere else the vaccine should be identical to the virus (and so we'll use the SA variant virus' spike amino there).

On the codon level, the codon of the existing Pfizer vaccine should be used, unless the SA variant has a different amino than the vaccine, in which case we should use a codon for the amino based on the LUT we created earlier from virus codons to vaccine codons (again, except for the 2 intentional changes of the vaccine vs the virus, where we will use the vaccine's codons as well).

```
vir_codons_that_need_encoding = []

for start_idx in range(len(vaccines_data[company_names[0]]['Amino Seq'])):
    vac_part = vaccines_data[company_names[0]]['Amino Seq'][start_idx]
    vir_part = modified_sa_vaccine_base_amino[start_idx]

    if vac_part!= vir_part and vir_part != ' ':
        vir_codons_that_need_encoding.append(modified_sa_vaccine_base_codon[start_idx * 3:(start_idx + 1)* 3])

vir_codons_that_need_encoding
```

```
['GCU', 'GGU', 'AAU', 'AAA', 'UAU', 'GGU', 'GUA', 'AAA', 'GUU']
```

```
for vir_codons_that_need_encoding in vir_codons_that_need_encoding:
    print(vir_codons_that_need_encoding, '-->', vir_to_vac_codons_map[vir_codons_that_need_encoding])
```

```
GCU --> defaultdict(<class 'int'>, {'GCC': 35, 'GCU': 5, 'GCA': 2})
GGU --> defaultdict(<class 'int'>, {'GGC': 40, 'GGG': 1, 'GGA': 5, 'GGU': 1})
AAU --> defaultdict(<class 'int'>, {'AAC': 41, 'AAU': 13})
AAA --> defaultdict(<class 'int'>, {'AAG': 31, 'AAA': 6})
UAU --> defaultdict(<class 'int'>, {'UAC': 36, 'UAU': 4})
GGU --> defaultdict(<class 'int'>, {'GGC': 40, 'GGG': 1, 'GGA': 5, 'GGU': 1})
GUA --> defaultdict(<class 'int'>, {'GUG': 15})
AAA --> defaultdict(<class 'int'>, {'AAG': 31, 'AAA': 6})
GUU --> defaultdict(<class 'int'>, {'GUG': 41, 'GUC': 6})
```

```
# This is the mapping we will use (when needed):
sa_to_vac_map = {
    'GCU': 'GCC',

    # We have this one twice, but that's still not enough to select different target codons (according to the distribution above
    'GGU': 'GGC',

    'AAU': 'AAC',

    # We have this one twice as well, but same as above
    'AAA': 'AAG',

    'UAU': 'UAC',
    'GUA': 'GUG',
    'GUU': 'GUG'
}
```

```
sa_vaccine = ""

for start_idx in range(len(vaccines_data[company_names[0]]['Amino Seq'])):
    vac_part = vaccines_data[company_names[0]]['Amino Seq'][start_idx]
    wuh_part = wuhan_sars_cov_2_data['Amino Seq'][start_idx]
    sav_part = modified_sa_vaccine_base_amino[start_idx]

    if wuh_part != vac_part:
        sa_vaccine += vaccines_data[company_names[0]]['Nucleotide Seq'][start_idx * 3:(start_idx + 1)* 3]
```

```
    else:
        if vac_part == sav_part:
            sa_vaccine += vaccines_data[company_names[0]]['Nucleotide Seq'][start_idx * 3:(start_idx + 1)* 3]
        else:
            if sav_part == ' ':
                continue
            else:
                sa_vaccine += sa_to_vac_map[modified_sa_vaccine_base_codon[start_idx * 3:(start_idx + 1)* 3]]
```

Let's check the GC content of the new vaccine's spike protein:

```
f'New vaccine (spike protein) GC content is: {GC(sa_vaccine):.2f}%.'
```

```
'New vaccine (spike protein) GC content is: 57.06%.'
```

The GC content is higher than the vaccine, so we will lower it by changing some of the codons in the mapping, as allowed by the mapping.

We tried lowering GC content by 3 nucleotides, but that resulted in a GC content of 56.98% which is too low.

We settled on changing the GC content by only 2 nucleotides, making the GC content of the new vaccine, slightly higher than that of the Pfizer vaccine (57.01%), working under the assumption that if the GC content can't be kept - slightly higher is better than slightly lower.

```
# This is the mapping we will use (when needed):
sa_to_vac_map = {
       # This is allowed by the mapping we have, so we won't increase GC content
       'GCU': 'GCU',

    # We have this one twice, but that's still not enough to select different target codons (according to the distribution above
    'GGU': 'GGC',

       # When we tried to not change this as well (also allowed by the mapping we have), we got a GC content of 56.98%.
       'AAU': 'AAC',

    # We have this one twice as well, but same as above
    'AAA': 'AAG',

       # This is also allowed by the mapping we have, so we won't increase GC content again
      'UAU': 'UAU',

    'GUA': 'GUG',
    'GUU': 'GUG'
}
```

```
sa_vaccine = ""

for start_idx in range(len(vaccines_data[company_names[0]]['Amino Seq'])):
    vac_part = vaccines_data[company_names[0]]['Amino Seq'][start_idx]
    wuh_part = wuhan_sars_cov_2_data['Amino Seq'][start_idx]
    sav_part = modified_sa_vaccine_base_amino[start_idx]

    if wuh_part != vac_part:
        sa_vaccine += vaccines_data[company_names[0]]['Nucleotide Seq'][start_idx * 3:(start_idx + 1)* 3]
    else:
        if vac_part == sav_part:
            sa_vaccine += vaccines_data[company_names[0]]['Nucleotide Seq'][start_idx * 3:(start_idx + 1)* 3]
        else:
            if sav_part == ' ':
                continue
            else:
                sa_vaccine += sa_to_vac_map[modified_sa_vaccine_base_codon[start_idx * 3:(start_idx + 1)* 3]]


f'New vaccine (spike protein) GC content is: {GC(sa_vaccine):.2f}%.'
```

```
'New vaccine (spike protein) GC content is: 57.01%.'
```

Checking that the new SA variant vaccine spike has the same structure as the SA variant virus on the amino level.

```
translated_sa_vaccine_spike = Seq.Seq(sa_vaccine).translate()
vaccine_base_amino_no_spaces = modified_sa_vaccine_base_amino.replace(' ', '')

print('Vac: New vaccine based on SA variant  -  Vir: SA variant virus\n============================================================')

step = 68
for start_idx in range(0, len(vaccines_data[company_names[0]]['Amino Seq']), step):
    vac_part = translated_sa_vaccine_spike[start_idx:start_idx + step]
    vir_part = vaccine_base_amino_no_spaces[start_idx:start_idx + step]

    print('Vac:', vac_part)
    print('Vir:', vir_part)
    print('    ', "".join([' ' if vac_part[i] == vir_part[i] else '^' for i in range(len(vac_part))]))
```

```
Vac: New vaccine based on SA variant  -  Vir: SA variant virus
============================================================

Vac: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI
Vir: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI

Vac: HVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
Vir: HVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC

Vac: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY
```

```
Vir: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY

Vac: SKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTF
Vir: SKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTF

Vac: LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFN
Vir: LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFN

Vac: ATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIA
Vir: ATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIA

Vac: PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTP
Vir: PGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTP

Vac: CNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLT
Vir: CNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLT

Vac: GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGV
Vir: GTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGV

Vac: NCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRR
Vir: NCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRR

Vac: ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSN
Vir: ARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSN

Vac: LLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIE
Vir: LLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIE

Vac: DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWT
Vir: DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWT

Vac: FGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQN
Vir: FGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQN

Vac: AQALNTLVKQLSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASAN
Vir: AQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASAN
                                    ^^
Vac: LAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPR
Vir: LAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPR

Vac: EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNH
Vir: EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNH

Vac: TSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAI
Vir: TSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAI

Vac: VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
Vir: VMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

Compare the spikes of the Pfizer vaccine and the new SA variant vaccine we created (we want to see the same vaccine, except for the differences that stem from the differences between the two virus variants):

```
inconsistency_start = modified_sa_vaccine_base_amino.find('HRSYLTPGDSSSGWTAGAAAYYVGYLQPRTF')

translated_sa_vaccine_spike = Seq.Seq(sa_vaccine).translate()
translated_modified_sa_vaccine_spike = f"{translated_sa_vaccine_spike[:inconsistency_start - 3]}   {translated_sa_vaccine_spike

print('VacS: New vaccine based on SA variant  -  VacP: Pfizer vaccine\n============================================================

step = 68
for start_idx in range(0, len(vaccines_data[company_names[0]]['Amino Seq']), step):
    vacs_part = translated_modified_sa_vaccine_spike[start_idx:start_idx + step]
    vacp_part = vaccines_data[company_names[0]]['Amino Seq'][start_idx:start_idx + step]

    print('VacS:', vacs_part)
    print('VacP:', vacp_part)
    print('     ', "".join([' ' if vacs_part[i] == vacp_part[i] else '^' for i in range(len(vacs_part))]))
```

```
VacS: New vaccine based on SA variant  -  VacP: Pfizer vaccine
============================================================

VacS: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI
VacP: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI

VacS: HVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
VacP: HVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
                  ^
VacS: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY
VacP: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY

VacS: SKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTL   HRSYLTPGDSSSGWTAGAAAYYVGYLQP
VacP: SKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP
                ^                           ^^^
VacS: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE
VacP: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE

VacS: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR
VacP: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR

VacS: QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
VacP: QIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
              ^
VacS: STPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
VacP: STPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
              ^               ^
VacS: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY
VacP: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY

VacS: QGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
VacP: QDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
       ^
VacS: PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
VacP: PRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
                        ^
VacS: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS
VacP: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS

VacS: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS
VacP: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS
```

```
VacS: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV
VacP: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV

VacS: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
VacP: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA

VacS: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH
VacP: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH

VacS: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF
VacP: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF

VacS: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL
VacP: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL

VacS: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VacP: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

```
# Add missing parts to the SA vaccine (end codon will be taken together with the 3' UTR from the Pfizer vaccine):
sa_vac_final = f"{vaccines_mRNA_seq[company_names[0]][:vaccines_data[company_names[0]]['Start Codon Index']]}{sa_vaccine}{vaccii
```

Compare the complete Pfizer vaccine to the new one for the SA variant on the nucleotide level:

```
inconsistency_start = sa_vac_final.find('ACAGAAGCUACCUGACACCUGGCGAUAGCAGCAGCGGA')

modified_sa_vac_final = f"{sa_vac_final[:inconsistency_start]}         {sa_vac_final[inconsistency_start:]}"

print('VacS: New vaccine based on SA variant  -  VacP: Pfizer vaccine\n============================================================
=

step = 102
for start_idx in range(0, len(sa_vac_final), step):
    vacs_part = modified_sa_vac_final[start_idx:start_idx + step]
    vacp_part = vaccines_mRNA_seq[company_names[0]][start_idx:start_idx + step]

    print('VacS:', vacs_part)
    print('VacP:', vacp_part)
    print('     ', "".join([' ' if vacs_part[i] == vacp_part[i] else '^' for i in range(len(vacs_part))]))
```

```
VacS: New vaccine based on SA variant  -  VacP: Pfizer vaccine
================================================================

VacS: GAGAAUAAACUAGUAUUCUUCUGGUCCCCACAGACUCAGAGAGAACCCGCCACCAUGUUCGUGUUCCUGGUGCUGCUGCCUCUGGUGUCCAGCCAGUGUGUG
VacP: GAGAAUAAACUAGUAUUCUUCUGGUCCCCACAGACUCAGAGAGAACCCGCCACCAUGUUCGUGUUCCUGGUGCUGCUGCCUCUGGUGUCCAGCCAGUGUGUG

VacS: AACCUGACCACCAGAACACAGCUGCCUCCAGCCUACACCAACAGCUUUACCAGAGGCGUGUACUACCCCGACAAGGUGUUCAGAUCCAGCGUGCUGCACUCU
VacP: AACCUGACCACCAGAACACAGCUGCCUCCAGCCUACACCAACAGCUUUACCAGAGGCGUGUACUACCCCGACAAGGUGUUCAGAUCCAGCGUGCUGCACUCU

VacS: ACCCAGGACCUGUUCCUGCCUUUCUUCAGCAACGUGACCUGGUUCCACGCCAUCCACGUGUCCGGCACCAAUGGCACCAAGAGAUUCGCUAACCCCGUGCUG
VacP: ACCCAGGACCUGUUCCUGCCUUUCUUCAGCAACGUGACCUGGUUCCACGCCAUCCACGUGUCCGGCACCAAUGGCACCAAGAGAUUCGACAACCCCGUGCUG
                                                                                                    ^^
```

```
VacS: CCCUUCAACGACGGGGUGUACUUUGCCAGCACCGAGAAGUCCAACAUCAUCAGAGGCUGGAUCUUCGGCACCACACUGGACAGCAAGACCCAGAGCCUGCUG
VacP: CCCUUCAACGACGGGGUGUACUUUGCCAGCACCGAGAAGUCCAACAUCAUCAGAGGCUGGAUCUUCGGCACCACACUGGACAGCAAGACCCAGAGCCUGCUG

VacS: AUCGUGAACAACGCCACCAACGUGGUCAUCAAAGUGUGCGAGUUCCAGUUCUGCAACGACCCCUUCCUGGGCGUCUACUACCACAAGAACAACAAGAGCUGG
VacP: AUCGUGAACAACGCCACCAACGUGGUCAUCAAAGUGUGCGAGUUCCAGUUCUGCAACGACCCCUUCCUGGGCGUCUACUACCACAAGAACAACAAGAGCUGG

VacS: AUGGAAAGCGAGUUCCGGGUGUACAGCAGCGCCAACAACUGCACCUUCGAGUACGUGUCCCAGCCUUUCCUGAUGGACCUGGAAGGCAAGCAGGGCAACUUC
VacP: AUGGAAAGCGAGUUCCGGGUGUACAGCAGCGCCAACAACUGCACCUUCGAGUACGUGUCCCAGCCUUUCCUGAUGGACCUGGAAGGCAAGCAGGGCAACUUC

VacS: AAGAACCUGCGCGAGUUCGUGUUUAAGAACAUCGACGGCUACUUCAAGAUCUACAGCAAGCACACCCCUAUCAACCUCGUGCGGGGCCUGCCUCAGGGCUUC
VacP: AAGAACCUGCGCGAGUUCGUGUUUAAGAACAUCGACGGCUACUUCAAGAUCUACAGCAAGCACACCCCUAUCAACCUCGUGCGGGAUCUGCCUCAGGGCUUC
                                                                                                        ^^

VacS: UCUGCUCUGGAACCCCUGGUGGAUCUGCCCAUCGGCAUCAACAUCACCCGGUUUCAGACACUGC          ACAGAAGCUACCUGACACCUGGCGAUAGC
VacP: UCUGCUCUGGAACCCCUGGUGGAUCUGCCCAUCGGCAUCAACAUCACCCGGUUUCAGACACUGCUGGCCCUGCACAGAAGCUACCUGACACCUGGCGAUAGC
                                                                    ^^^^^^^^^

VacS: AGCAGCGGAUGGACAGCUGGUGCCGCCGCUUACUAUGUGGGCUACCUGCAGCCUAGAACCUUCCUGCUGAAGUACAACGAGAACGGCACCAUCACCGACGCC
VacP: AGCAGCGGAUGGACAGCUGGUGCCGCCGCUUACUAUGUGGGCUACCUGCAGCCUAGAACCUUCCUGCUGAAGUACAACGAGAACGGCACCAUCACCGACGCC

VacS: GUGGAUUGUGCUCUGGAUCCUCUGAGCGAGACAAAGUGCACCCUGAAGUCCUUCACCGUGGAAAAGGGCAUCUACCAGACCAGCAACUUCCGGGUGCAGCCC
VacP: GUGGAUUGUGCUCUGGAUCCUCUGAGCGAGACAAAGUGCACCCUGAAGUCCUUCACCGUGGAAAAGGGCAUCUACCAGACCAGCAACUUCCGGGUGCAGCCC

VacS: ACCGAAUCCAUCGUGCGGUUCCCCAAUAUCACCAAUCUGUGCCCCUUCGGCGAGGUGUUCAAUGCCACCAGAUUCGCCUCUGUGUACGCCUGGAACCGGAAG
VacP: ACCGAAUCCAUCGUGCGGUUCCCCAAUAUCACCAAUCUGUGCCCCUUCGGCGAGGUGUUCAAUGCCACCAGAUUCGCCUCUGUGUACGCCUGGAACCGGAAG

VacS: CGGAUCAGCAAUUGCGUGGCCGACUACUCCGUGCUGUACAACUCCGCCAGCUUCAGCACCUUCAAGUGCUACGGCGUGUCCCCUACCAAGCUGAACGACCUG
VacP: CGGAUCAGCAAUUGCGUGGCCGACUACUCCGUGCUGUACAACUCCGCCAGCUUCAGCACCUUCAAGUGCUACGGCGUGUCCCCUACCAAGCUGAACGACCUG

VacS: UGCUUCACAAACGUGUACGCCGACAGCUUCGUGAUCCGGGGAGAUGAAGUGCGGCAGAUUGCCCCUGGACAGACAGGCAACAUCGCCGACUACAACUACAAG
VacP: UGCUUCACAAACGUGUACGCCGACAGCUUCGUGAUCCGGGGAGAUGAAGUGCGGCAGAUUGCCCCUGGACAGACAGGCAAGAUCGCCGACUACAACUACAAG
                                                                                                ^

VacS: CUGCCCGACGACUUCACCGGCUGUGUGAUUGCCUGGAACAGCAACAACCUGGACUCCAAAGUCGGCGGCAACUACAAUUACCUGUACCGGCUGUUCCGGAAG
VacP: CUGCCCGACGACUUCACCGGCUGUGUGAUUGCCUGGAACAGCAACAACCUGGACUCCAAAGUCGGCGGCAACUACAAUUACCUGUACCGGCUGUUCCGGAAG

VacS: UCCAAUCUGAAGCCCUUCGAGCGGGACAUCUCCACCGAGAUCUAUCAGGCCGGCAGCACCCCUUGUAACGGCGUGAAGGGCUUCAACUGCUACUUCCCACUG
VacP: UCCAAUCUGAAGCCCUUCGAGCGGGACAUCUCCACCGAGAUCUAUCAGGCCGGCAGCACCCCUUGUAACGGCGUGGAAGGCUUCAACUGCUACUUCCCACUG
                                                                                              ^ ^

VacS: CAGUCCUACGGCUUUCAGCCCACAUAUGGCGUGGGCUAUCAGCCCUACAGAGUGGUGGUGCUGAGCUUCGAACUGCUGCAUGCCCCUGCCACAGUGUGCGGC
VacP: CAGUCCUACGGCUUUCAGCCCACAAAUGGCGUGGGCUAUCAGCCCUACAGAGUGGUGGUGCUGAGCUUCGAACUGCUGCAUGCCCCUGCCACAGUGUGCGGC
                           ^

VacS: CCUAAGAAAAGCACCAAUCUCGUGAAGAACAAAUGCGUGAACUUCAACUUCAACGGCCUGACCGGCACCGGCGUGCUGACAGAGAGCAACAAGAAGUUCCUG
VacP: CCUAAGAAAAGCACCAAUCUCGUGAAGAACAAAUGCGUGAACUUCAACUUCAACGGCCUGACCGGCACCGGCGUGCUGACAGAGAGCAACAAGAAGUUCCUG

VacS: CCAUUCCAGCAGUUUGGCCGGGAUAUCGCCGAUACCACAGACGCCGUUAGAGAUCCCCAGACACUGGAAAUCCUGGACAUCACCCCUUGCAGCUUCGGCGGA
VacP: CCAUUCCAGCAGUUUGGCCGGGAUAUCGCCGAUACCACAGACGCCGUUAGAGAUCCCCAGACACUGGAAAUCCUGGACAUCACCCCUUGCAGCUUCGGCGGA

VacS: GUGUCUGUGAUCACCCCUGGCACCAACACCAGCAAUCAGGUGGCAGUGCUGUACCAGGGCGUGAACUGUACCGAAGUGCCCGUGGCCAUUCACGCCGAUCAG
VacP: GUGUCUGUGAUCACCCCUGGCACCAACACCAGCAAUCAGGUGGCAGUGCUGUACCAGGACGUGAACUGUACCGAAGUGCCCGUGGCCAUUCACGCCGAUCAG
                                                                ^

VacS: CUGACACCUACAUGGCGGGUGUACUCCACCGGCAGCAAUGUGUUUCAGACCAGAGCCGGCUGUCUGAUCGGAGCCGAGCACGUGAACAAUAGCUACGAGUGC
VacP: CUGACACCUACAUGGCGGGUGUACUCCACCGGCAGCAAUGUGUUUCAGACCAGAGCCGGCUGUCUGAUCGGAGCCGAGCACGUGAACAAUAGCUACGAGUGC

VacS: GACAUCCCCAUCGGCGCUGGAAUCUGCGCCAGCUACCAGACACAGACAAACAGCCCUCGGAGAGCCAGAAGCGUGGCCAGCCAGAGCAUCAUUGCCUACACA
VacP: GACAUCCCCAUCGGCGCUGGAAUCUGCGCCAGCUACCAGACACAGACAAACAGCCCUCGGAGAGCCAGAAGCGUGGCCAGCCAGAGCAUCAUUGCCUACACA

VacS: AUGUCUCUGGGCGUGGAGAACAGCGUGGCCUACUCCAACAACUCUAUCGCUAUCCCCACCAACUUCACCAUCAGCGUGACCACAGAGAUCCUGCCUGUGUCC
VacP: AUGUCUCUGGGCGCCGAGAACAGCGUGGCCUACUCCAACAACUCUAUCGCUAUCCCCACCAACUUCACCAUCAGCGUGACCACAGAGAUCCUGCCUGUGUCC
                  ^^
VacS: AUGACCAAGACCAGCGUGGACUGCACCAUGUACAUCUGCGGCGAUUCCACCGAGUGCUCCAACCUGCUGCUGCAGUACGGCAGCUUCUGCACCCAGCUGAAU
VacP: AUGACCAAGACCAGCGUGGACUGCACCAUGUACAUCUGCGGCGAUUCCACCGAGUGCUCCAACCUGCUGCUGCAGUACGGCAGCUUCUGCACCCAGCUGAAU

VacS: AGAGCCCUGACAGGGAUCGCCGUGGAACAGGACAAGAACACCCAAGAGGUGUUCGCCCAAGUGAAGCAGAUCUACAAGACCCCUCCUAUCAAGGACUUCGGC
```

```
VacP: AGAGCCCUGACAGGGAUCGCCGUGGAACAGGACAAGAACACCCAAGAGGUGUUCGCCCAAGUGAAGCAGAUCUACAAGACCCCUCCUAUCAAGGACUUCGGC

VacS: GGCUUCAAUUUCAGCCAGAUUCUGCCCGAUCCUAGCAAGCCCAGCAAGCGGAGCUUCAUCGAGGACCUGCUGUUCAACAAAGUGACACUGGCCGACGCCGGC
VacP: GGCUUCAAUUUCAGCCAGAUUCUGCCCGAUCCUAGCAAGCCCAGCAAGCGGAGCUUCAUCGAGGACCUGCUGUUCAACAAAGUGACACUGGCCGACGCCGGC

VacS: UUCAUCAAGCAGUAUGGCGAUUGUCUGGGCGACAUUGCCGCCAGGGAUCUGAUUUGCGCCCAGAAGUUUAACGGACUGACAGUGCUGCCUCCUCUGCUGACC
VacP: UUCAUCAAGCAGUAUGGCGAUUGUCUGGGCGACAUUGCCGCCAGGGAUCUGAUUUGCGCCCAGAAGUUUAACGGACUGACAGUGCUGCCUCCUCUGCUGACC

VacS: GAUGAGAUGAUCGCCCAGUACACAUCUGCCCUGCUGGCCGGCACAAUCACAAGCGGCUGGACAUUUGGAGCAGGCGCCGCUCUGCAGAUCCCCUUUGCUAUG
VacP: GAUGAGAUGAUCGCCCAGUACACAUCUGCCCUGCUGGCCGGCACAAUCACAAGCGGCUGGACAUUUGGAGCAGGCGCCGCUCUGCAGAUCCCCUUUGCUAUG

VacS: CAGAUGGCCUACCGGUUCAACGGCAUCGGAGUGACCCAGAAUGUGCUGUACGAGAACCAGAAGCUGAUCGCCAACCAGUUCAACAGCGCCAUCGGCAAGAUC
VacP: CAGAUGGCCUACCGGUUCAACGGCAUCGGAGUGACCCAGAAUGUGCUGUACGAGAACCAGAAGCUGAUCGCCAACCAGUUCAACAGCGCCAUCGGCAAGAUC

VacS: CAGGACAGCCUGAGCAGCACAGCAAGCGCCCUGGGAAAGCUGCAGGACGUGGUCAACCAGAAUGCCCAGGCACUGAACACCCUGGUCAAGCAGCUGUCCUCC
VacP: CAGGACAGCCUGAGCAGCACAGCAAGCGCCCUGGGAAAGCUGCAGGACGUGGUCAACCAGAAUGCCCAGGCACUGAACACCCUGGUCAAGCAGCUGUCCUCC

VacS: AACUUCGGCGCCAUCAGCUCUGUGCUGAACGAUAUCCUGAGCAGACUGGACCCUCCUGAGGCCGAGGUGCAGAUCGACAGACUGAUCACAGGCAGACUGCAG
VacP: AACUUCGGCGCCAUCAGCUCUGUGCUGAACGAUAUCCUGAGCAGACUGGACCCUCCUGAGGCCGAGGUGCAGAUCGACAGACUGAUCACAGGCAGACUGCAG

VacS: AGCCUCCAGACAUACGUGACCCAGCAGCUGAUCAGAGCCGCCGAGAUUAGAGCCUCUGCCAAUCUGGCCGCCACCAAGAUGUCUGAGUGUGUGCUGGGCCAG
VacP: AGCCUCCAGACAUACGUGACCCAGCAGCUGAUCAGAGCCGCCGAGAUUAGAGCCUCUGCCAAUCUGGCCGCCACCAAGAUGUCUGAGUGUGUGCUGGGCCAG

VacS: AGCAAGAGAGUGGACUUUUGCGGCAAGGGCUACCACCUGAUGAGCUUCCCUCAGUCUGCCCCUCACGGCGUGGUGUUUCUGCACGUGACAUAUGUGCCCGCU
VacP: AGCAAGAGAGUGGACUUUUGCGGCAAGGGCUACCACCUGAUGAGCUUCCCUCAGUCUGCCCCUCACGGCGUGGUGUUUCUGCACGUGACAUAUGUGCCCGCU

VacS: CAAGAGAAGAAUUUCACCACCGCUCCAGCCAUCUGCCACGACGGCAAAGCCCACUUUCCUAGAGAAGGCGUGUUCGUGUCCAACGGCACCCAUUGGUUCGUG
VacP: CAAGAGAAGAAUUUCACCACCGCUCCAGCCAUCUGCCACGACGGCAAAGCCCACUUUCCUAGAGAAGGCGUGUUCGUGUCCAACGGCACCCAUUGGUUCGUG

VacS: ACACAGCGGAACUUCUACGAGCCCCAGAUCAUCACCACCGACAACACCUUCGUGUCUGGCAACUGCGACGUCGUGAUCGGCAUUGUGAACAAUACCGUGUAC
VacP: ACACAGCGGAACUUCUACGAGCCCCAGAUCAUCACCACCGACAACACCUUCGUGUCUGGCAACUGCGACGUCGUGAUCGGCAUUGUGAACAAUACCGUGUAC

VacS: GACCCUCUGCAGCCCGAGCUGGACAGCUUCAAAGAGGAACUGGACAAGUACUUUAAGAACCACACAAGCCCCGACGUGGACCUGGGCGAUAUCAGCGGAAUC
VacP: GACCCUCUGCAGCCCGAGCUGGACAGCUUCAAAGAGGAACUGGACAAGUACUUUAAGAACCACACAAGCCCCGACGUGGACCUGGGCGAUAUCAGCGGAAUC

VacS: AAUGCCAGCGUCGUGAACAUCCAGAAAGAGAUCGACCGGCUGAACGAGGUGGCCAAGAAUCUGAACGAGAGCCUGAUCGACCUGCAAGAACUGGGGAAGUAC
VacP: AAUGCCAGCGUCGUGAACAUCCAGAAAGAGAUCGACCGGCUGAACGAGGUGGCCAAGAAUCUGAACGAGAGCCUGAUCGACCUGCAAGAACUGGGGAAGUAC

VacS: GAGCAGUACAUCAAGUGGCCCUGGUACAUCUGGCUGGGCUUUAUCGCCGGACUGAUUGCCAUCGUGAUGGUCACAAUCAUGCUGUGUUGCAUGACCAGCUGC
VacP: GAGCAGUACAUCAAGUGGCCCUGGUACAUCUGGCUGGGCUUUAUCGCCGGACUGAUUGCCAUCGUGAUGGUCACAAUCAUGCUGUGUUGCAUGACCAGCUGC

VacS: UGUAGCUGCCUGAAGGGCUGUUGUAGCUGUGGCAGCUGCUGCAAGUUCGACGAGGACGAUUCUGAGCCCGUGCUGAAGGGCGUGAAACUGCACUACACAUGA
VacP: UGUAGCUGCCUGAAGGGCUGUUGUAGCUGUGGCAGCUGCUGCAAGUUCGACGAGGACGAUUCUGAGCCCGUGCUGAAGGGCGUGAAACUGCACUACACAUGA

VacS: UGACUCGAGCUGGUACUGCAUGCACGCAAUGCUAGCUGCCCCUUUCCCGUCCUGGGUACCCCGAGUCUCCCCCGACCUCGGGUCCCAGGUAUGCUCCCACCU
VacP: UGACUCGAGCUGGUACUGCAUGCACGCAAUGCUAGCUGCCCCUUUCCCGUCCUGGGUACCCCGAGUCUCCCCCGACCUCGGGUCCCAGGUAUGCUCCCACCU

VacS: CCACCUGCCCCACUCACCACCUCUGCUAGUUCCAGACACCUCCCAAGCACGCAGCAAUGCAGCUCAAAACGCUUAGCCUAGCCACACCCCCACGGGAAACAG
VacP: CCACCUGCCCCACUCACCACCUCUGCUAGUUCCAGACACCUCCCAAGCACGCAGCAAUGCAGCUCAAAACGCUUAGCCUAGCCACACCCCCACGGGAAACAG

VacS: CAGUGAUUAACCUUUAGCAAUAAACGAAAGUUUAACUAAGCUAUACUAACCCCAGGGUUGGUCAAUUUCGUGCCAGCCACACCCUGGAGCUAGCA
VacP: CAGUGAUUAACCUUUAGCAAUAAACGAAAGUUUAACUAAGCUAUACUAACCCCAGGGUUGGUCAAUUUCGUGCCAGCCACACCCUGGAGCUAGCA
```

The final answer, the vaccine on the nucleotide level is:

```
step = 136

for start_idx in range(0, len(sa_vac_final), step):
    print(sa_vac_final[start_idx:start_idx + step])
```

```
GAGAAUAAACUAGUAUUCUUCUGGUCCCCACAGACUCAGAGAGAACCCGCCACCAUGUUCGUGUUCCUGGUGCUGCUGCCUCUGGUGUCCAGCCAGUGUGUGAACCUGACCACCAGAACACAGCUGCCU
ACACCAACAGCUUUACCAGAGGCGUGUACUACCCCGACAAGGUGUUCAGAUCCAGCGUGCUGCACUCUACCCAGGACCUGUUCCUGCCUUUCUUCAGCAACGUGACCUGGUUCCACGCCAUCCACGUGU
CAAUGGCACCAAGAGAUUCGCUAACCCCGUGCUGCCCUUCAACGACGGGGUGUACUUUGCCAGCACCGAGAAGUCCAACAUCAUCAGAGGCUGGAUCUUCGGCACCACACUGGACAGCAAGACCCAGAC
AUCGUGAACAACGCCACCAACGUGGUCAUCAAAGUGUGCGAGUUCCAGUUCUGCAACGACCCCUUCCUGGGCGUCUACUACCACAAGAACAACAAGAGCUGGAUGGAAAGCGAGUUCCGGGUGUACAGC
ACAACUGCACCUUCGAGUACGUGUCCCAGCCUUUCCUGAUGGACCUGGAAGGCAAGCAGGGCAACUUCAAGAACCUGCGCGAGUUCGUGUUUAAGAACAUCGACGGCUACUUCAAGAUCUACAGCAAGC
UAUCAACCUCGUGCGGGGCCUGCCUCAGGGCUUCUCUGCUCUGGAACCCCUGGUGGAUCUGCCCAUCGGCAUCAACAUCACCCGGUUUCAGACACUGCACAGAAGCUACCUGACACCUGGCGAUAGCAC
UGGACAGCUGGUGCCGCCGCUUACUAUGUGGGCUACCUGCAGCCUAGAACCUUCCUGCUGAAGUACAACGAGAACGGCACCAUCACCGACGCCGUGGAUUGUGCUCUGGAUCCUCUGAGCGAGACAAAC
UGAAGUCCUUCACCGUGGAAAAGGGCAUCUACCAGACCAGCAACUUCCGGGUGCAGCCCACCGAAUCCAUCGUGCGGUUCCCCAAUAUCACCAAUCUGUGCCCCUUCGGCGAGGUGUUCAAUGCCACCA
CUCUGUGUACGCCUGGAACCGGAAGCGGAUCAGCAAUUGCGUGGCCGACUACUCCGUGCUGUACAACUCCGCCAGCUUCAGCACCUUCAAGUGCUACGGCGUGUCCCCUACCAAGCUGAACGACCUGUC
AACGUGUACGCCGACAGCUUCGUGAUCCGGGGAGAUGAAGUGCGGCAGAUUGCCCCUGGACAGACAGGCAACAUCGCCGACUACAACUACAAGCUGCCCGACGACUUCACCGGCUGUGUGAUUGCCUGC
ACAACCUGGACUCCAAAGUCGGCGGCAACUACAAUUACCUGUACCGGCUGUUCCGGAAGUCCAAUCUGAAGCCCUUCGAGCGGGACAUCUCCACCGAGAUCUAUCAGGCCGGCAGCACCCCUUGUAACC
GGGCUUCAACUGCUACUUCCCACUGCAGUCCUACGGCUUUCAGCCCACAUAUGGCGUGGGCUAUCAGCCCUACAGAGUGGUGGUGCUGAGCUUCGAACUGCUGCAUGCCCCUGCCACAGUGUGCGGCCC
AGCACCAAUCUCGUGAAGAACAAAUGCGUGAACUUCAACUUCAACGGCCUGACCGGCACCGGCGUGCUGACAGAGAGCAACAAGAAGUUCCUGCCAUUCCAGCAGUUUGGCCGGGAUAUCGCCGAUACC
CCGUUAGAGAUCCCCAGACACUGGAAAUCCUGGACAUCACCCCUUGCAGCUUCGGCGGAGUGUCUGUGAUCACCCCUGGCACCAACACCAGCAAUCAGGUGGCAGUGCUGUACCAGGGCGUGAACUGUA
GCCCGUGGCCAUUCACGCCGAUCAGCUGACACCUACAUGGCGGGUGUACUCCACCGGCAGCAAUGUGUUUCAGACCAGAGCCGGCUGUCUGAUCGGAGCCGAGCACGUGAACAAUAGCUACGAGUGCGA
AUCGGCGCUGGAAUCUGCGCCAGCUACCAGACACAGACAAACAGCCCUCGGAGAGCCAGAAGCGUGGCCAGCCAGAGCAUCAUUGCCUACACAAUGUCUCUGGGCGUGGAGAACAGCGUGGCCUACUCC
CUAUCGCUACUCCCCACCAACUUCACCAUCAGCGUGACCACAGAGAUCCUGCCUGUGUCCAUGACCAAGACCAGCGUGGACUGCACCAUGUACAUCUGCGGCGAUUCCACCGAGUGCUCCAACCUGCUGC
CGGCAGCUUCUGCACCCAGCUGAAUAGAGCCCUGACAGGGAUCGCCGUGGAACAGGACAAGAACACCCAAGAGGUGUUCGCCCAAGUGAAGCAGAUCUACAAGACCCCUCCUAUCAAGGACUUCGGCGC
UUCAGCCAGAUUCUGCCCGAUCCUAGCAAGCCCAGCAAGCGGAGCUUCAUCGAGGACCUGCUGUUCAACAAAGUGACACUGGCCGACGCCGGCUUCAUCAAGCAGUAUGGCGAUUGUCUGGGCGACAUC
GGGAUCUGAUUUGCGCCCAGAAGUUUAACGGACUGACAGUGCUGCCUCCUCUGCUGACCGAUGAGAUGAUCGCCCAGUACACAUCUGCCCUGCUGGCCGGCACAAUCACAAGCGGCUGGACAUUUGGAC
CGCUCUGCAGAUCCCCUUUGCUAUGCAGAUGGCCUACCGGUUCAACGGCAUCGGAGUGACCCAGAAUGUGCUGUACGAGAACCAGAAGCUGAUCGCCAACCAGUUCAACAGCGCCAUCGGCAAGAUCCA
CUGAGCAGCACAGCAAGCGCCCUGGGAAAGCUGCAGGACGUGGUCAACCAGAAUGCCCAGGCACUGAACACCCUGGUCAAGCAGCUGUCCUCCAACUUCGGCGCCAUCAGCUCUGUGCUGAACGAUAUC
GACUGGACCCUCCUGAGGCCGAGGUGCAGAUCGACAGACUGAUCACAGGCAGACUGCAGAGCCUCCAGACAUACGUGACCCAGCAGCUGAUCAGAGCCGCCGAGAUUAGAGCCUCUGCCAAUCUGGCCC
GAUGUCUGAGUGUGUGCUGGGCCAGAGCAAGAGAGUGGACUUUUGCGGCAAGGGCUACCACCUGAUGAGCUUCCCUCAGUCUGCCCCUCACGGCGUGGUGUUUCUGCACGUGACAUAUGUGCCCGCUCA
AAUUUCACCACCGCUCCAGCCAUCUGCCACGACGGCAAAGCCCACUUUCCUAGAGAAGGCGUGUUCGUGUCCAACGGCACCCAUUGGUUCGUGACACAGCGGAACUUCUACGAGCCCCAGAUCAUCACC
ACACCUUCGUGUCUGGCAACUGCGACGUCGUGAUCGGCAUUGUGAACAAUACCGUGUACGACCCUCUGCAGCCCGAGCUGGACAGCUUCAAAGAGGAACUGGACAAGUACUUUAAGAACCACACAAGCC
GGACCUGGGCGAUAUCAGCGGAAUCAAUGCCAGCGUCGUGAACAUCCAGAAAGAGAUCGACCGGCUGAACGAGGUGGCCAAGAAUCUGAACGAGAGCCUGAUCGACCUGCAAGAACUGGGGAAGUACGA
AUCAAGUGGCCCUGGUACAUCUGGCUGGGCUUUAUCGCCGGACUGAUUGCCAUCGUGAUGGUCACAAUCAUGCUGUUGUCAUGACCAGCUGCUGUAGCUGCCUGAAGGGCUGUUGUAGCUGUGGCAGC
AGUUCGACGAGGACGAUUCUGAGCCCGUGCUGAAGGGCGUGAAACUGCACUACACAUGAUGACUCGAGCUGGUACUGCAUGCACGCAAUGCUAGCUGCCCCUUUCCCGUCCUGGGUACCCCGAGUCUCC
UCGGGUCCCAGGUAUGCUCCCACCUCCACCUGCCCCACUCACCACCUCUGCUAGUUCCAGACACCUCCCAAGCACGCAGCAAUGCAGCUCAAAACGCUUAGCCUAGCCACACCCCCACGGGAAACAGCA
ACCUUUAGCAAUAAACGAAAGUUUAACUAAGCUAUACUAACCCCAGGGUUGGUCAAUUUCGUGCCAGCCACACCCUGGAGCUAGCA
```

(a possibly) Contradictory finding

We also found a paper by SA researchers (et al.): https://www.medrxiv.org/content/10.1101/2020.12.21.20248640v1.full-text , that specifically discusses the spike protein mutations relevant to the SA variant.

The researchers discuss a total of 9 mutations, the most distinguishing of which is **E484K** (as it was found present in less than 0.02% of the samples outside of SA).

They also divide the 9 mutations into 2 groups:

- 6 that were present since the first sampling of the SA variant on 2020-10-15: **D614G**, **D80A**, **D215G**, **E484K**, **N501Y**, **A701V**.
- 3 that appeared by end of 2020-11: **L18F**, **R246I**, **K417N**.

Furthermore, the **LAL** sequence we found to be missing in our classification of the proteins is also mentioned in the paper (as **L242_244L**), however it is still debated due to a different mutation (at least at the time of the paper's publishing - 2020-12-22).

We explore the samples and the specified mutations.

Checking if there are any samples that contain all the mutations:

```
mutations = ['D614G', 'D80A', 'D215G', 'E484K', 'N501Y', 'A701V', 'L18F', 'R246I', 'K417N']

mutation_sites = [int(m[1:-1]) - 1 for m in mutations]
mutation_values = [m[-1:] for m in mutations]

[k for k, d in sa_spike_proteins.items() if all([d['Amino Seq'][mutation_site] == mutation_values[mutation_idx] for mutation_id
```

```
[]
```

No, but let's look at the samples, and see if they have sequncing ambiguities in the mutation sites:

```
mutations = ['D614G', 'D80A', 'D215G', 'E484K', 'N501Y', 'A701V', 'L18F', 'R246I', 'K417N']
mutation_sites = [int(m[1:-1]) - 1 for m in mutations]
mutation_values = [m[-1:] for m in mutations]
original_values = [m[:1] for m in mutations]

print('Original', original_values)
print('Expected', mutation_values)
[(k, [d['Amino Seq'][mutation_site] for mutation_site in mutation_sites]) for k, d in sa_spike_proteins.items()]
```

```
Original ['D', 'D', 'D', 'E', 'N', 'A', 'L', 'R', 'K']
Expected ['G', 'A', 'G', 'K', 'Y', 'V', 'F', 'I', 'N']
```

```
[('MW912490.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW913437.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW914008.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW914015.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW914445.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW914542.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW905790.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW905844.1', ['C', 'A', 'G', 'N', 'G', 'S', 'F', 'L', 'D']),
 ('MW905875.1', ['C', 'A', 'G', 'N', 'G', 'S', 'F', 'L', 'D']),
 ('MW906028.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW906061.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW907199.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW907323.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW908243.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW908815.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW909222.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW909347.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW898265.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW883160.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW877177.1', ['C', 'A', 'G', 'N', 'G', 'S', 'F', 'L', 'D']),
 ('MW880890.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW869153.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW844243.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW844258.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW844743.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW847393.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW849825.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW795351.1', ['C', 'A', 'G', 'N', 'G', 'S', 'F', 'L', 'D']),
 ('MW842292.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW822592.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('HG999935.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('HG999939.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW807936.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW808030.1', ['C', 'A', 'G', 'N', 'G', 'S', 'F', 'L', 'D']),
 ('MW808712.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW809047.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW803575.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW803586.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW793569.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW796654.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW796843.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW789246.1', ['C', 'A', 'G', 'N', 'G', 'S', 'F', 'L', 'D']),
 ('MW792684.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW792756.1', ['C', 'A', 'G', 'N', 'G', 'S', 'F', 'L', 'D']),
 ('MW793005.1', ['C', 'A', 'G', 'N', 'G', 'S', 'F', 'L', 'D']),
 ('MW773815.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW763124.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763125.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763126.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763127.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763128.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763129.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'X', 'N']),
 ('MW763130.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763131.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763132.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW687146.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW580244.1', ['C', 'A', 'G', 'N', 'G', 'S', 'F', 'L', 'D']),
 ('MW617734.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW621453.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW598408.1', ['C', 'A', 'G', 'N', 'G', 'S', 'F', 'L', 'D']),
 ('MW598413.1', ['C', 'A', 'G', 'N', 'G', 'S', 'F', 'L', 'D']),
```

```
 ('MW598419.1', ['C', 'A', 'G', 'N', 'G', 'S', 'L', 'L', 'D']),
 ('MW580574.1', ['C', 'A', 'G', 'N', 'G', 'S', 'F', 'L', 'D']),
 ('MW580576.1', ['C', 'A', 'G', 'N', 'G', 'S', 'F', 'L', 'D'])]
```

```
mutations = ['D614G', 'D80A', 'D215G', 'E484K', 'N501Y', 'A701V', 'L18F', 'R246I', 'K417N']
mutation_sites = [int(m[1:-1]) - 1 for m in mutations]
mutation_values = [m[-1:] for m in mutations]
original_values = [m[:1] for m in mutations]

print('Original', original_values)
print('Expected', mutation_values)
[(k, [d['Amino Seq'][mutation_site] for mutation_site in mutation_sites]) for k, d in sa_spike_proteins.items() if d['Amino Seq
```

```
Original ['D', 'D', 'D', 'E', 'N', 'A', 'L', 'R', 'K']
Expected ['G', 'A', 'G', 'K', 'Y', 'V', 'F', 'I', 'N']

[('MW763124.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763125.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763126.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763127.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763128.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763129.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'X', 'N']),
 ('MW763130.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763131.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N']),
 ('MW763132.1', ['G', 'A', 'G', 'K', 'Y', 'V', 'L', 'R', 'N'])]
```

We see that:

- Only 9 samples have the highly distinguishing **E484K** mutation.
- The same 9 samples also have all 6 "early" mutations, and only 1 of the "later" mutations.
- There's only one sample with sequencing ambiguities, and it is contained within the same 9 samples.

Let's have a look at the similarity between the proteins of these samples:

```
sa_samples = [k for k, d in sa_spike_proteins.items() if d['Amino Seq'][mutation_sites[3]] == mutation_values[3]]
sa_proteins = [d['Amino Seq'] for k, d in sa_spike_proteins.items() if d['Amino Seq'][mutation_sites[3]] == mutation_values[3]]
```

```
s1 = sa_proteins[0]

step = 136
for start_idx in range(0, len(s1), step):
    for s in sa_proteins:
        print(s[start_idx:start_idx + step])
```

```
    s_parts = [s[start_idx:start_idx + step] for s in sa_proteins]
    s_is = [set([s_part[i] for s_part in s_parts]) for i in range(len(s_parts[0]))]
    comparison = [' ' if len(s_i) == 1 else '^' for s_i in s_is]
    print("".join(comparison))
```

```
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVI

NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTXXXLHRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFXXXXXXHRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFXXXXXLHRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFXXXXXLHRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTXXXXXXSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTXXXLHRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFXXXXXLHRSYLTPGDSSSGWTAGAAA
NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTXXXXXRSYLTPGDSSSGWTAGAAA
                                                                                                      ^^^^^^^^
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF
RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSF

QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN
QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKN

GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA
GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICA

PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
```

```
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD
PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD

FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA
FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASA

NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA

FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYI

IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

1. It appears all proteins are identical, except for sequencing ambiguities.
2. Completely contradictory to our analysis, all samples either have the **LAL** amino sequence, or are as long as the Wuhan protein sequence (in which case the area covering the **LAL** sequence has at least partial sequencing ambiguities).
3. The only sequence without ambiguities is the first one.

```
sa_vaccine_base_paper = sa_proteins[0]
```

And again use majority vote to decide what the nucleotide level is:

```
sa_vaccine_base_nts_paper = ""

for amino_idx, _ in enumerate(sa_vaccine_base_paper):
    codons_for_amino = defaultdict(int)

    for protein in sa_proteins:
        if protein[amino_idx] not in 'XBZJ':
            for sample_id_with_protein in unique_sa_spike_proteins[protein]:
                codon_for_amino_in_sample = sa_spike_proteins[sample_id_with_protein]['Nucleotide Seq'][amino_idx * 3:(amino_id
                codons_for_amino[codon_for_amino_in_sample] += 1

    selected_codon = max(codons_for_amino, key=codons_for_amino.get)
    sa_vaccine_base_nts_paper += selected_codon
```

Basing a vaccine on this sequence however, is bound to give different results than what we found so far.

Let's explore this in the same way we created the previous vaccine:

```
print('Vir1: Previously found SA variant virus  -  Vir2: SA variant virus accoring to paper\n==================================
step = 68
for start_idx in range(0, len(vaccines_data[company_names[0]]['Amino Seq']), step):

    vir1_part = modified_sa_vaccine_base_amino[start_idx:start_idx + step]
    vir2_part = sa_vaccine_base_paper[start_idx:start_idx + step]

    print('Vir1:', vir1_part)
    print('Vir2:', vir2_part)
    print('     ', "".join([' ' if vir1_part[i] == vir2_part[i] else '^' for i in range(len(vir1_part))]))
```

```
Vir1: Previously found SA variant virus  -  Vir2: SA variant virus accoring to paper
================================================================================

Vir1: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI
Vir2: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI

Vir1: HVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
Vir2: HVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC

Vir1: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY
Vir2: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY
```

```
Vir1: SKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTL   HRSYLTPGDSSSGWTAGAAAYYVGYLQP
Vir2: SKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP
                                           ^^^
Vir1: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE
Vir2: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE

Vir1: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR
Vir2: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR

Vir1: QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
Vir2: QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG

Vir1: STPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
Vir2: STPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN

Vir1: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY
Vir2: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY

Vir1: QGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
Vir2: QGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS

Vir1: PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
Vir2: PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE

Vir1: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS
Vir2: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS

Vir1: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS
Vir2: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS

Vir1: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV
Vir2: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV

Vir1: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
Vir2: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA

Vir1: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH
Vir2: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH

Vir1: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF
Vir2: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF

Vir1: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL
Vir2: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL

Vir1: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
Vir2: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

Surprisingly, the spikes are identical on the amino level, except for the extra **LAL** sequence we used in the new vaccine.

Let's compare the nucleotide level:

```
print('VirPap: New spike based on paper  -  VirPrv: Previous spike we found\n==========================================================')

step = 102
for start_idx in range(0, len(modified_sa_vaccine_base_codon), step):
    virprv_part = modified_sa_vaccine_base_codon[start_idx:start_idx + step]
    virpap_part = sa_vaccine_base_nts_paper[start_idx:start_idx + step]

    print('VirPrv:', virprv_part)
    print('VirPap:', virpap_part)
    print('       ', "".join([' ' if virprv_part[i] == virpap_part[i] else '^' for i in range(len(virprv_part))]))
```

```
VirPap: New spike based on paper  -  VirPrv: Previous spike we found
====================================================================

VirPrv: AUGUUUGUUUUUCUUGUUUUAUUGCCACUAGUCUCUAGUCAGUGUGUUAAUCUUACAACCAGAACUCAAUUACCCCCUGCAUACACUAAUUCUUUCACACGU
VirPap: AUGUUUGUUUUUCUUGUUUUAUUGCCACUAGUCUCUAGUCAGUGUGUUAAUCUUACAACCAGAACUCAAUUACCCCCUGCAUACACUAAUUCUUUCACACGU

VirPrv: GGUGUUUAUUACCCUGACAAAGUUUUCAGAUCCUCAGUUUUACAUUCAACUCAGGACUUGUUCUUACCUUUCUUUUCCAAUGUUACUUGGUUCCAUGCUAUA
VirPap: GGUGUUUAUUACCCUGACAAAGUUUUCAGAUCCUCAGUUUUACAUUCAACUCAGGACUUGUUCUUACCUUUCUUUUCCAAUGUUACUUGGUUCCAUGCUAUA

VirPrv: CAUGUCUCUGGGACCAAUGGUACUAAGAGGUUUGCUAACCCUGUCCUACCAUUUAAUGAUGGUGUUUAUUUUGCUUCCACUGAGAAGUCUAACAUAAUAAGA
VirPap: CAUGUCUCUGGGACCAAUGGUACUAAGAGGUUUGCUAACCCUGUCCUACCAUUUAAUGAUGGUGUUUAUUUUGCUUCCACUGAGAAGUCUAACAUAAUAAGA

VirPrv: GGCUGGAUUUUUGGUACUACUUUAGAUUCGAAGACCCAGUCCCUACUUAUUGUUAAUAACGCUACUAAUGUUGUUAUUAAAGUCUGUGAAUUUCAAUUUUGU
VirPap: GGCUGGAUUUUUGGUACUACUUUAGAUUCGAAGACCCAGUCCCUACUUAUUGUUAAUAACGCUACUAAUGUUGUUAUUAAAGUCUGUGAAUUUCAAUUUUGU

VirPrv: AAUGAUCCAUUUUUGGGUGUUUAUUACCACAAAAACAACAAAAGUUGGAUGGAAAGUGAGUUCAGAGUUUAUUCUAGUGCGAAUAAUUGCACUUUUGAAUAU
VirPap: AAUGAUCCAUUUUUGGGUGUUUAUUACCACAAAAACAACAAAAGUUGGAUGGAAAGUGAGUUCAGAGUUUAUUCUAGUGCGAAUAAUUGCACUUUUGAAUAU

VirPrv: GUCUCUCAGCCUUUUCUUAUGGACCUUGAAGGAAAACAGGGUAAUUUCAAAAAUCUUAGGGAAUUUGUGUUUAAGAAUAUUGAUGGUUAUUUUAAAAUAUAU
VirPap: GUCUCUCAGCCUUUUCUUAUGGACCUUGAAGGAAAACAGGGUAAUUUCAAAAAUCUUAGGGAAUUUGUGUUUAAGAAUAUUGAUGGUUAUUUUAAAAUAUAU

VirPrv: UCUAAGCACACGCCUAUUAAUUUAGUGCGUGGUCUCCCUCAGGGUUUUUCGGCUUUAGAACCAUUGGUAGAUUUGCCAAUAGGUAUUAACAUCACUAGGUUU
VirPap: UCUAAGCACACGCCUAUUAAUUUAGUGCGUGGUCUCCCUCAGGGUUUUUCGGCUUUAGAACCAUUGGUAGAUUUGCCAAUAGGUAUUAACAUCACUAGGUUU

VirPrv: CAAACUUUA          CAUAGAAGUUAUUUGACUCCUGGUGAUUCUUCUUCAGGUUGGACAGCUGGUGCUGCAGCUUAUUAUGUGGGUUAUCUUCAACCU
VirPap: CAAACUUUACUUGCUUUACAUAGAAGUUAUUUGACUCCUGGUGAUUCUUCUUCAGGUUGGACAGCUGGUGCUGCAGCUUAUUAUGUGGGUUAUCUUCAACCU
                  ^^^^^^^^^
VirPrv: AGGACUUUUCUAUUAAAAUAUAAUGAAAAUGGAACCAUUACAGAUGCUGUAGACUGUGCACUUGACCCUCUCUCAGAAACAAAGUGUACGUUGAAAUCCUUC
VirPap: AGGACUUUUCUAUUAAAAUAUAAUGAAAAUGGAACCAUUACAGAUGCUGUAGACUGUGCACUUGACCCUCUCUCAGAAACAAAGUGUACGUUGAAAUCCUUC

VirPrv: ACUGUAGAAAAAGGAAUCUAUCAAACUUCUAACUUUAGAGUCCAACCAACAGAAUCUAUUGUUAGAUUUCCUAAUAUUACAAACUUGUGCCCUUUUGGUGAA
VirPap: ACUGUAGAAAAAGGAAUCUAUCAAACUUCUAACUUUAGAGUCCAACCAACAGAAUCUAUUGUUAGAUUUCCUAAUAUUACAAACUUGUGCCCUUUUGGUGAA

VirPrv: GUUUUUAACGCCACCAGAUUUGCAUCUGUUUAUGCUUGGAACAGGAAGAGAAUCAGCAACUGUGUUGCUGAUUAUUCUGUCCUAUAUAAUUCCGCAUCAUUU
VirPap: GUUUUUAACGCCACCAGAUUUGCAUCUGUUUAUGCUUGGAACAGGAAGAGAAUCAGCAACUGUGUUGCUGAUUAUUCUGUCCUAUAUAAUUCCGCAUCAUUU

VirPrv: UCCACUUUUAAGUGUUAUGGAGUGUCUCCUACUAAAUUAAAUGAUCUCUGCUUUACUAAUGUCUAUGCAGAUUCAUUUGUAAUUAGAGGUGAUGAAGUCAGA
VirPap: UCCACUUUUAAGUGUUAUGGAGUGUCUCCUACUAAAUUAAAUGAUCUCUGCUUUACUAAUGUCUAUGCAGAUUCAUUUGUAAUUAGAGGUGAUGAAGUCAGA

VirPrv: CAAAUCGCUCCAGGGCAAACUGGAAAUAUUGCUGAUUAUAAUUAUAAAUUACCAGAUGAUUUUACAGGCUGCGUUAUAGCUUGGAAUUCUAACAAUCUUGAU
VirPap: CAAAUCGCUCCAGGGCAAACUGGAAAUAUUGCUGAUUAUAAUUAUAAAUUACCAGAUGAUUUUACAGGCUGCGUUAUAGCUUGGAAUUCUAACAAUCUUGAU

VirPrv: UCUAAGGUUGGUGGUAAUUAUAAUUACCUGUAUAGAUUGUUUAGGAAGUCUAAUCUCAAACCUUUUGAGAGAGAUAUUUCAACUGAAAUCUAUCAGGCCGGU
VirPap: UCUAAGGUUGGUGGUAAUUAUAAUUACCUGUAUAGAUUGUUUAGGAAGUCUAAUCUCAAACCUUUUGAGAGAGAUAUUUCAACUGAAAUCUAUCAGGCCGGU

VirPrv: AGCACACCUUGUAAUGGUGUUAAAGGUUUUAAUUGUUACUUUCCUUUACAAUCAUAUGGUUUCCAACCCACUUAUGGUGUUGGUUACCAACCAUACAGAGUA
VirPap: AGCACACCUUGUAAUGGUGUUAAAGGUUUUAAUUGUUACUUUCCUUUACAAUCAUAUGGUUUCCAACCCACUUAUGGUGUUGGUUACCAACCAUACAGAGUA
```

```
VirPrv: GUAGUACUUUCUUUUGAACUUCUACAUGCACCAGCAACUGUUUGUGGACCUAAAAAGUCUACUAAUUUGGUUAAAAACAAAUGUGUCAAUUUCAACUUCAAU
VirPap: GUAGUACUUUCUUUUGAACUUCUACAUGCACCAGCAACUGUUUGUGGACCUAAAAAGUCUACUAAUUUGGUUAAAAACAAAUGUGUCAAUUUCAACUUCAAU

VirPrv: GGUUUAACAGGCACAGGUGUUCUUACUGAGUCUAACAAAAAGUUUCUGCCUUUCCAACAAUUUGGCAGAGACAUUGCUGACACUACUGAUGCUGUCCGUGAU
VirPap: GGUCUAACAGGCACAGGUGUUCUUACUGAGUCUAACAAAAAGUUUCUGCCUUUCCAACAAUUUGGCAGAGACAUUGCUGACACUACUGAUGCUGUCCGUGAU
           ^
VirPrv: CCACAGACACUUGAGAUUCUUGACAUUACACCAUGUUCUUUUGGUGGUGUCAGUGUUAUAACACCAGGAACAAAUACUUCUAACCAGGUUGCUGUUCUUUAU
VirPap: CCACAGACACUUGAGAUUCUUGACAUUACACCAUGUUCUUUUGGUGGUGUCAGUGUUAUAACACCAGGAACAAAUACUUCUAACCAGGUUGCUGUUCUUUAU

VirPrv: CAGGGUGUUAACUGCACAGAAGUCCCUGUUGCUAUUCAUGCAGAUCAACUUACUCCUACUUGGCGUGUUUAUUCUACAGGUUCUAAUGUUUUUCAAACACGU
VirPap: CAGGGUGUUAACUGCACAGAAGUCCCUGUUGCUAUUCAUGCAGAUCAACUUACUCCUACUUGGCGUGUUUAUUCUACAGGUUCUAAUGUUUUUCAAACACGU

VirPrv: GCAGGCUGUUUAAUAGGGGCUGAACAUGUCAACAACUCAUAUGAGUGUGACAUACCCAUUGGUGCAGGUAUAUGCGCUAGUUAUCAGACUCAGACUAAUUCU
VirPap: GCAGGCUGUUUAAUAGGGGCUGAACAUGUCAACAACUCAUAUGAGUGUGACAUACCCAUUGGUGCAGGUAUAUGCGCUAGUUAUCAGACUCAGACUAAUUCU

VirPrv: CCUCGGCGGGCACGUAGUGUAGCUAGUCAAUCCAUCAUUGCCUACACUAUGUCACUUGGUGUAGAAAAUUCAGUUGCUUACUCUAAUAACUCUAUUGCCAUA
VirPap: CCUCGGCGGGCACGUAGUGUAGCUAGUCAAUCCAUCAUUGCCUACACUAUGUCACUUGGUGUAGAAAAUUCAGUUGCUUACUCUAAUAACUCUAUUGCCAUA

VirPrv: CCCACAAAUUUUACUAUUAGUGUUACCACAGAAAUUCUACCAGUGUCUAUGACCAAGACAUCAGUAGAUUGUACAAUGUACAUUUGUGGUGAUUCAACUGAA
VirPap: CCCACAAAUUUUACUAUUAGUGUUACCACAGAAAUUCUACCAGUGUCUAUGACCAAGACAUCAGUAGAUUGUACAAUGUACAUUUGUGGUGAUUCAACUGAA

VirPrv: UGCAGCAAUCUUUUGUUGCAAUAUGGCAGUUUUUGUACACAAUUAAACCGUGCUUUAACUGGAAUAGCUGUUGAACAAGACAAAAACACCCAAGAAGUUUUU
VirPap: UGCAGCAAUCUUUUGUUGCAAUAUGGCAGUUUUUGUACACAAUUAAACCGUGCUUUAACUGGAAUAGCUGUUGAACAAGACAAAAACACCCAAGAAGUUUUU

VirPrv: GCACAAGUCAAACAAAUUUACAAAACACCACCAAUUAAAGAUUUUGGUGGUUUUAAUUUUUCACAAAUAUUACCAGAUCCAUCAAAACCAAGCAAGAGGUCA
VirPap: GCACAAGUCAAACAAAUUUACAAAACACCACCAAUUAAAGAUUUUGGUGGUUUUAAUUUUUCACAAAUAUUACCAGAUCCAUCAAAACCAAGCAAGAGGUCA

VirPrv: UUUAUUGAAGAUCUACUUUUCAACAAAGUGACACUUGCAGAUGCUGGCUUCAUCAAACAAUAUGGUGAUUGCCUUGGUGAUAUUGCUGCUAGAGACCUCAUU
VirPap: UUUAUUGAAGAUCUACUUUUCAACAAAGUGACACUUGCAGAUGCUGGCUUCAUCAAACAAUAUGGUGAUUGCCUUGGUGAUAUUGCUGCUAGAGACCUCAUU

VirPrv: UGUGCACAAAAGUUUAACGGCCUUACUGUUUUGCCACCUUUGCUCACAGAUGAAAUGAUUGCUCAAUACACUUCUGCACUGUUAGCGGGUACAAUCACUUCU
VirPap: UGUGCACAAAAGUUUAACGGCCUUACUGUUUUGCCACCUUUGCUCACAGAUGAAAUGAUUGCUCAAUACACUUCUGCACUGUUAGCGGGUACAAUCACUUCU

VirPrv: GGUUGGACCUUUGGUGCAGGUGCUGCAUUACAAAUACCAUUUGCUAUGCAAAUGGCUUAUAGGUUUAAUGGUAUUGGAGUUACACAGAAUGUUCUCUAUGAG
VirPap: GGUUGGACCUUUGGUGCAGGUGCUGCAUUACAAAUACCAUUUGCUAUGCAAAUGGCUUAUAGGUUUAAUGGUAUUGGAGUUACACAGAAUGUUCUCUAUGAG

VirPrv: AACCAAAAAUUGAUUGCCAACCAAUUUAAUAGUGCUAUUGGCAAAAUUCAAGACUCACUUUCUUCCACAGCAAGUGCACUUGGAAAACUUCAAGAUGUGGUC
VirPap: AACCAAAAAUUGAUUGCCAACCAAUUUAAUAGUGCUAUUGGCAAAAUUCAAGACUCACUUUCUUCCACAGCAAGUGCACUUGGAAAACUUCAAGAUGUGGUC

VirPrv: AACCAAAAUGCACAAGCUUUAAACACGCUUGUUAAACAACUUAGCUCCAAUUUUGGUGCAAUUUCAAGUGUUUUAAAUGAUAUCCUUUCACGUCUUGACAAA
VirPap: AACCAAAAUGCACAAGCUUUAAACACGCUUGUUAAACAACUUAGCUCCAAUUUUGGUGCAAUUUCAAGUGUUUUAAAUGAUAUCCUUUCACGUCUUGACAAA

VirPrv: GUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCAGAAAUCAGAGCU
VirPap: GUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUAUGUGACUCAACAAUUAAUUAGAGCUGCAGAAAUCAGAGCU

VirPrv: UCUGCUAAUCUUGCUGCUACUAAAAUGUCAGAGUGUGUACUUGGACAAUCAAAAAGAGUUGAUUUUUGUGGAAAGGGCUAUCAUCUUAUGUCCUUCCCUCAG
VirPap: UCUGCUAAUCUUGCUGCUACUAAAAUGUCAGAGUGUGUACUUGGACAAUCAAAAAGAGUUGAUUUUUGUGGAAAGGGCUAUCAUCUUAUGUCCUUCCCUCAG

VirPrv: UCAGCACCUCAUGGUGUAGUCUUCUUGCAUGUGACUUAUGUCCCUGCACAAGAAAAGAACUUCACAACUGCUCCUGCCAUUUGUCAUGAUGGAAAAGCACAC
VirPap: UCAGCACCUCAUGGUGUAGUCUUCUUGCAUGUGACUUAUGUCCCUGCACAAGAAAAGAACUUCACAACUGCUCCUGCCAUUUGUCAUGAUGGAAAAGCACAC

VirPrv: UUUCCUCGUGAAGGUGUCUUUGUUUCAAAUGGCACACACUGGUUUGUAACACAAAGGAAUUUUUAUGAACCACAAAUCAUUACUACAGACAACACAUUUGUG
VirPap: UUUCCUCGUGAAGGUGUCUUUGUUUCAAAUGGCACACACUGGUUUGUAACACAAAGGAAUUUUUAUGAACCACAAAUCAUUACUACAGACAACACAUUUGUG

VirPrv: UCUGGUAACUGUGAUGUUGUAAUAGGAAUUGUCAACAACACAGUUUAUGAUCCUUUGCAACCUGAAUUAGACUCAUUCAAGGAGGAGUUAGAUAAAUAUUUU
VirPap: UCUGGUAACUGUGAUGUUGUAAUAGGAAUUGUCAACAACACAGUUUAUGAUCCUUUGCAACCUGAAUUAGACUCAUUCAAGGAGGAGUUAGAUAAAUAUUUU

VirPrv: AAGAAUCAUACAUCACCAGAUGUUGAUUUAGGUGACAUCUCUGGCAUUAAUGCUUCAGUUGUAAACAUUCAAAAAGAAAUUGACCGCCUCAAUGAGGUUGCC
VirPap: AAGAAUCAUACAUCACCAGAUGUUGAUUUAGGUGACAUCUCUGGCAUUAAUGCUUCAGUUGUAAACAUUCAAAAAGAAAUUGACCGCCUCAAUGAGGUUGCC
```

```
VirPrv: AAGAAUUUAAAUGAAUCUCUCAUCGAUCUCCAAGAACUUGGAAAGUAUGAGCAGUAUAUAAAAUGGCCAUGGUACAUUUGGCUAGGUUUUAUAGCUGGCUUG
VirPap: AAGAAUUUAAAUGAAUCUCUCAUCGAUCUCCAAGAACUUGGAAAGUAUGAGCAGUAUAUAAAAUGGCCAUGGUACAUUUGGCUAGGUUUUAUAGCUGGCUUG

VirPrv: AUUGCCAUAGUAAUGGUGACAAUUAUGCUUUGCUGUAUGACCAGUUGCUGUAGUUGUCUCAAGGGCUGUUGUUCUUGUGGAUCCUGCUGCAAAUUUGAUGAA
VirPap: AUUGCCAUAGUAAUGGUGACAAUUAUGCUUUGCUGUAUGACCAGUUGCUGUAGUUGUCUCAAGGGCUGUUGUUCUUGUGGAUCCUGCUGCAAAUUUGAUGAA

VirPrv: GACGACUCUGAGCCAGUGCUCAAAGGAGUCAAAUUACAUUACACA
VirPap: GACGACUCUGAGCCAGUGCUCAAAGGAGUCAAAUUACAUUACACA
```

Only one nucleotide's difference between the spike according to the paper, and the method we previously used.

And compare to the Pfizer vaccine:

```
print('Vac: Pfizer vaccine  -  Vir: SA variant virus\n==========================================================================\n')

step = 68
for start_idx in range(0, len(vaccines_data[company_names[0]]['Amino Seq']), step):

    vac_part = vaccines_data[company_names[0]]['Amino Seq'][start_idx:start_idx + step]
    vir_part = sa_vaccine_base_paper[start_idx:start_idx + step]

    print('Vac:', vac_part)
    print('Vir:', vir_part)
    print('    ', "".join([' ' if vac_part[i] == vir_part[i] else '^' for i in range(len(vac_part))]))
```

```
Vac: Pfizer vaccine  -  Vir: SA variant virus
==========================================================================

Vac: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI
Vir: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI

Vac: HVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
Vir: HVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
                ^
Vac: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY
Vir: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY

Vac: SKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP
Vir: SKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP
               ^
Vac: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE
Vir: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE

Vac: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR
Vir: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR

Vac: QIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
Vir: QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
```

```
            ^
Vac: STPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
Vir: STPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
           ^                ^
Vac: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY
Vir: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY

Vac: QDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
Vir: QGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
      ^
Vac: PRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
Vir: PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
                         ^
Vac: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS
Vir: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS

Vac: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS
Vir: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS

Vac: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV
Vir: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV

Vac: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
Vir: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
                                       ^^
Vac: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH
Vir: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH

Vac: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF
Vir: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF

Vac: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL
Vir: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL

Vac: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
Vir: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

Codons that need encoding are the same as before:

```
vir_codons_that_need_encoding = []

for start_idx in range(len(vaccines_data[company_names[0]]['Amino Seq'])):
    vac_part = vaccines_data[company_names[0]]['Amino Seq'][start_idx]
    vir_part = sa_vaccine_base_paper[start_idx]

    if vac_part != vir_part:
        vir_codons_that_need_encoding.append(sa_vaccine_base_nts_paper[start_idx * 3:(start_idx + 1)* 3])

vir_codons_that_need_encoding
```

```
['GCU', 'GGU', 'AAU', 'AAA', 'UAU', 'GGU', 'GUA', 'AAA', 'GUU']
```

```
for vir_codons_that_need_encoding in vir_codons_that_need_encoding:
    print(vir_codons_that_need_encoding, '-->', vir_to_vac_codons_map[vir_codons_that_need_encoding])
```

```
GCU --> defaultdict(<class 'int'>, {'GCC': 35, 'GCU': 5, 'GCA': 2})
GGU --> defaultdict(<class 'int'>, {'GGC': 40, 'GGG': 1, 'GGA': 5, 'GGU': 1})
AAU --> defaultdict(<class 'int'>, {'AAC': 41, 'AAU': 13})
AAA --> defaultdict(<class 'int'>, {'AAG': 31, 'AAA': 6})
UAU --> defaultdict(<class 'int'>, {'UAC': 36, 'UAU': 4})
GGU --> defaultdict(<class 'int'>, {'GGC': 40, 'GGG': 1, 'GGA': 5, 'GGU': 1})
GUA --> defaultdict(<class 'int'>, {'GUG': 15})
AAA --> defaultdict(<class 'int'>, {'AAG': 31, 'AAA': 6})
GUU --> defaultdict(<class 'int'>, {'GUG': 41, 'GUC': 6})
```

```
# This is the mapping we will use (when needed):
sa_to_vac_map = {
    'GCU': 'GCC',

    # We have this one twice
    'GGU': 'GGC',

    'AAU': 'AAC',

    # We have this one twice as well
    'AAA': 'AAG',

    'UAU': 'UAC',
    'GUA': 'GUG',
    'GUU': 'GUG'
}
```

```
sa_vaccine_paper = ""

for start_idx in range(len(vaccines_data[company_names[0]]['Amino Seq'])):
    vac_part = vaccines_data[company_names[0]]['Amino Seq'][start_idx]
    wuh_part = wuhan_sars_cov_2_data['Amino Seq'][start_idx]
    sav_part = sa_vaccine_base_paper[start_idx]

    if wuh_part != vac_part:
        sa_vaccine_paper += vaccines_data[company_names[0]]['Nucleotide Seq'][start_idx * 3:(start_idx + 1)* 3]
    else:
        if vac_part == sav_part:
            sa_vaccine_paper += vaccines_data[company_names[0]]['Nucleotide Seq'][start_idx * 3:(start_idx + 1)* 3]
        else:
            sa_vaccine_paper += sa_to_vac_map[sa_spike_proteins[sa_samples[0]]['Nucleotide Seq'][start_idx * 3:(start_idx + 1)*
```

Let's check the GC content of the new vaccine's spike protein:

```
f'New vaccine (spike protein) GC content is: {GC(sa_vaccine_paper):.2f}%.'
```

```
'New vaccine (spike protein) GC content is: 57.11%.'
```

Again, the GC is too high, so we repeat the process of not encoding all codons to those with higher GC content:

```
sa_to_vac_map = {
       # This is allowed by the mapping we have, so we won't increase GC content
       'GCU': 'GCU',

       # This is allowed by the mapping we have, so we won't increase GC content
       # We have this one twice
       'GGU': 'GGA',

    # When we tried to not change this as well, we again got a GC content of 56.98%.
    'AAU': 'AAC',

    # We have this one twice as well, but same as above
    'AAA': 'AAG',

       # This is also allowed by the mapping we have, so we won't increase GC content again
      'UAU': 'UAU',

    'GUA': 'GUG',
    'GUU': 'GUG'
}


sa_vaccine_paper = ""

for start_idx in range(len(vaccines_data[company_names[0]]['Amino Seq'])):
    vac_part = vaccines_data[company_names[0]]['Amino Seq'][start_idx]
    wuh_part = wuhan_sars_cov_2_data['Amino Seq'][start_idx]
    sav_part = sa_vaccine_base_paper[start_idx]

    if wuh_part != vac_part:
        sa_vaccine_paper += vaccines_data[company_names[0]]['Nucleotide Seq'][start_idx * 3:(start_idx + 1)* 3]
    else:
        if vac_part == sav_part:
            sa_vaccine_paper += vaccines_data[company_names[0]]['Nucleotide Seq'][start_idx * 3:(start_idx + 1)* 3]
        else:
            sa_vaccine_paper += sa_to_vac_map[sa_spike_proteins[sa_samples[0]]['Nucleotide Seq'][start_idx * 3:(start_idx + 1)*



f'New vaccine (spike protein) GC content is: {GC(sa_vaccine_paper):.2f}%.'
```

```
'New vaccine (spike protein) GC content is: 57.00%.'
```

And again as before, we had to select between either having a GC content of 57%, or 56.98%, and chose the higher one.

Checking that the new SA variant vaccine spike has the same structure as the SA variant virus on the amino level.

```
translated_sa_vaccine_spike_paper = Seq.Seq(sa_vaccine_paper).translate()

print('Vac: New vaccine based on SA variant  -  Vir: SA variant virus\n============================================================')

step = 68
for start_idx in range(0, len(vaccines_data[company_names[0]]['Amino Seq']), step):
    vac_part = translated_sa_vaccine_spike_paper[start_idx:start_idx + step]
    vir_part = sa_vaccine_base_paper[start_idx:start_idx + step]

    print('Vac:', vac_part)
    print('Vir:', vir_part)
    print('    ', "".join([' ' if vac_part[i] == vir_part[i] else '^' for i in range(len(vac_part))]))
```

```
Vac: New vaccine based on SA variant  -  Vir: SA variant virus
============================================================

Vac: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI
Vir: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI

Vac: HVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
Vir: HVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC

Vac: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY
Vir: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY

Vac: SKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP
Vir: SKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP

Vac: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE
Vir: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE

Vac: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR
Vir: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR

Vac: QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
Vir: QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG

Vac: STPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
Vir: STPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN

Vac: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY
Vir: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY

Vac: QGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
Vir: QGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS

Vac: PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
Vir: PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE

Vac: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS
Vir: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS
```

```
Vac: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS
Vir: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS

Vac: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV
Vir: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV

Vac: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
Vir: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
                                       ^^
Vac: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH
Vir: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH

Vac: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF
Vir: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF

Vac: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL
Vir: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL

Vac: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
Vir: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

Compare the spikes of the Pfizer vaccine and the new SA variant vaccine we created (we want to see the same vaccine, except for the differences that stem from the differences between the two virus variants):

```
print('VacS: New vaccine based on SA variant  -  VacP: Pfizer vaccine\n============================================================')

step = 68
for start_idx in range(0, len(vaccines_data[company_names[0]]['Amino Seq']), step):
    vacs_part = translated_sa_vaccine_spike_paper[start_idx:start_idx + step]
    vacp_part = vaccines_data[company_names[0]]['Amino Seq'][start_idx:start_idx + step]

    print('VacS:', vacs_part)
    print('VacP:', vacp_part)
    print('     ', "".join([' ' if vacs_part[i] == vacp_part[i] else '^' for i in range(len(vacs_part))]))
```

```
VacS: New vaccine based on SA variant  -  VacP: Pfizer vaccine
============================================================

VacS: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI
VacP: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI

VacS: HVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
VacP: HVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
                ^
VacS: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY
VacP: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY

VacS: SKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP
VacP: SKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP
                ^
VacS: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE
VacP: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE
```

```
VacS: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR
VacP: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR

VacS: QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
VacP: QIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
             ^
VacS: STPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
VacP: STPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
             ^                ^
VacS: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY
VacP: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY

VacS: QGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
VacP: QDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
       ^
VacS: PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
VacP: PRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
                          ^
VacS: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS
VacP: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS

VacS: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS
VacP: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS

VacS: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV
VacP: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV

VacS: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
VacP: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA

VacS: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH
VacP: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH

VacS: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF
VacP: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF

VacS: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL
VacP: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL

VacS: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VacP: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

Compare the new SA variant vaccine based on the paper, and the one we previously found:

```
print('VacPap: New vaccine based on paper  -  VacPrv: Previous vaccine we found\n==============================================')

step = 68
for start_idx in range(0, len(translated_modified_sa_vaccine_spike), step):
    vacpap_part = translated_sa_vaccine_spike_paper[start_idx:start_idx + step]
    vacprv_part = translated_modified_sa_vaccine_spike[start_idx:start_idx + step]

    print('VacPap:', vacpap_part)
```

```
    print('VacPrv:', vacprv_part)
    print('       ', "".join([' ' if vacpap_part[i] == vacprv_part[i] else '^' for i in range(len(vacpap_part))]))
```

```
VacPap: New vaccine based on paper  -  VacPrv: Previous vaccine we found
======================================================================

VacPap: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI
VacPrv: MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAI

VacPap: HVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC
VacPrv: HVSGTNGTKRFANPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFC

VacPap: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY
VacPrv: NDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIY

VacPap: SKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQP
VacPrv: SKHTPINLVRGLPQGFSALEPLVDLPIGINITRFQTL   HRSYLTPGDSSSGWTAGAAAYYVGYLQP
                                          ^^^
VacPap: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE
VacPrv: RTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE

VacPap: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR
VacPrv: VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVR

VacPap: QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG
VacPrv: QIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG

VacPap: STPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
VacPrv: STPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN

VacPap: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY
VacPrv: GLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLY

VacPap: QGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
VacPrv: QGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS

VacPap: PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE
VacPrv: PRRARSVASQSIIAYTMSLGVENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE

VacPap: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS
VacPrv: CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS

VacPap: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS
VacPrv: FIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITS

VacPap: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV
VacPrv: GWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV

VacPap: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
VacPrv: NQNAQALNTLVKQLSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA

VacPap: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH
VacPrv: SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH

VacPap: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF
VacPrv: FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYF

VacPap: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL
VacPrv: KNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGL
```

```
VacPap: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
VacPrv: IAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

Unsurprisingly, the vaccines are identical on the amino level, except for the extra **LAL** sequence we used in the new vaccine.

Let's compare the nucleotide level:

```
inconsistency_start = sa_vaccine.find('ACAGAAGCUACCUGACACCUGGCGAUAGCAGCAGCGGA')

modified_sa_vac = f"{sa_vaccine[:inconsistency_start]}         {sa_vaccine[inconsistency_start:]}"

print('VacPap: New vaccine based on paper  -  VacPrv: Previous vaccine we found\n==========================================================')

step = 102
for start_idx in range(0, len(modified_sa_vac), step):
    vacprv_part = modified_sa_vac[start_idx:start_idx + step]
    vacpap_part = sa_vaccine_paper[start_idx:start_idx + step]

    print('VacPrv:', vacprv_part)
    print('VacPap:', vacpap_part)
    print('       ', "".join([' ' if vacprv_part[i] == vacpap_part[i] else '^' for i in range(len(vacprv_part))]))
```

```
VacPap: New vaccine based on paper  -  VacPrv: Previous vaccine we found
======================================================================

VacPrv: AUGUUCGUGUUCCUGGUGCUGCUGCCUCUGGUGUCCAGCCAGUGUGUGAACCUGACCACCAGAACACAGCUGCCUCCAGCCUACACCAACAGCUUUACCAGA
VacPap: AUGUUCGUGUUCCUGGUGCUGCUGCCUCUGGUGUCCAGCCAGUGUGUGAACCUGACCACCAGAACACAGCUGCCUCCAGCCUACACCAACAGCUUUACCAGA

VacPrv: GGCGUGUACUACCCCGACAAGGUGUUCAGAUCCAGCGUGCUGCACUCUACCCAGGACCUGUUCCUGCCUUUCUUCAGCAACGUGACCUGGUUCCACGCCAUC
VacPap: GGCGUGUACUACCCCGACAAGGUGUUCAGAUCCAGCGUGCUGCACUCUACCCAGGACCUGUUCCUGCCUUUCUUCAGCAACGUGACCUGGUUCCACGCCAUC

VacPrv: CACGUGUCCGGCACCAAUGGCACCAAGAGAUUCGCUAACCCCGUGCUGCCCUUCAACGACGGGGUGUACUUUGCCAGCACCGAGAAGUCCAACAUCAUCAGA
VacPap: CACGUGUCCGGCACCAAUGGCACCAAGAGAUUCGCUAACCCCGUGCUGCCCUUCAACGACGGGGUGUACUUUGCCAGCACCGAGAAGUCCAACAUCAUCAGA

VacPrv: GGCUGGAUCUUCGGCACCACACUGGACAGCAAGACCCAGAGCCUGCUGAUCGUGAACAACGCCACCAACGUGGUCAUCAAAGUGUGCGAGUUCCAGUUCUGC
VacPap: GGCUGGAUCUUCGGCACCACACUGGACAGCAAGACCCAGAGCCUGCUGAUCGUGAACAACGCCACCAACGUGGUCAUCAAAGUGUGCGAGUUCCAGUUCUGC

VacPrv: AACGACCCCUUCCUGGGCGUCUACUACCACAAGAACAACAAGAGCUGGAUGGAAAGCGAGUUCCGGGUGUACAGCAGCGCCAACAACUGCACCUUCGAGUAC
VacPap: AACGACCCCUUCCUGGGCGUCUACUACCACAAGAACAACAAGAGCUGGAUGGAAAGCGAGUUCCGGGUGUACAGCAGCGCCAACAACUGCACCUUCGAGUAC

VacPrv: GUGUCCCAGCCUUUCCUGAUGGACCUGGAAGGCAAGCAGGGCAACUUCAAGAACCUGCGCGAGUUCGUGUUUAAGAACAUCGACGGCUACUUCAAGAUCUAC
VacPap: GUGUCCCAGCCUUUCCUGAUGGACCUGGAAGGCAAGCAGGGCAACUUCAAGAACCUGCGCGAGUUCGUGUUUAAGAACAUCGACGGCUACUUCAAGAUCUAC

VacPrv: AGCAAGCACACCCCUAUCAACCUCGUGCGGGGCCUGCCUCAGGGCUUCUCUGCUCUGGAACCCCUGGUGGAUCUGCCCAUCGGCAUCAACAUCACCCGGUUU
VacPap: AGCAAGCACACCCCUAUCAACCUCGUGCGGGGACUGCCUCAGGGCUUCUCUGCUCUGGAACCCCUGGUGGAUCUGCCCAUCGGCAUCAACAUCACCCGGUUU
                                        ^
VacPrv: CAGACACUGC         ACAGAAGCUACCUGACACCUGGCGAUAGCAGCAGCGGAUGGACAGCUGGUGCCGCCGCUUACUAUGUGGGCUACCUGCAGCCU
VacPap: CAGACACUGCUGGCCCUGCACAGAAGCUACCUGACACCUGGCGAUAGCAGCAGCGGAUGGACAGCUGGUGCCGCCGCUUACUAUGUGGGCUACCUGCAGCCU
```

```
                 ^^^^^^^^^
VacPrv: AGAACCUUCCUGCUGAAGUACAACGAGAACGGCACCAUCACCGACGCCGUGGAUUGUGCUCUGGAUCCUCUGAGCGAGACAAAGUGCACCCUGAAGUCCUUC
VacPap: AGAACCUUCCUGCUGAAGUACAACGAGAACGGCACCAUCACCGACGCCGUGGAUUGUGCUCUGGAUCCUCUGAGCGAGACAAAGUGCACCCUGAAGUCCUUC

VacPrv: ACCGUGGAAAAGGGCAUCUACCAGACCAGCAACUUCCGGGUGCAGCCCACCGAAUCCAUCGUGCGGUUCCCCAAUAUCACCAAUCUGUGCCCCUUCGGCGAG
VacPap: ACCGUGGAAAAGGGCAUCUACCAGACCAGCAACUUCCGGGUGCAGCCCACCGAAUCCAUCGUGCGGUUCCCCAAUAUCACCAAUCUGUGCCCCUUCGGCGAG

VacPrv: GUGUUCAAUGCCACCAGAUUCGCCUCUGUGUACGCCUGGAACCGGAAGCGGAUCAGCAAUUGCGUGGCCGACUACUCCGUGCUGUACAACUCCGCCAGCUUC
VacPap: GUGUUCAAUGCCACCAGAUUCGCCUCUGUGUACGCCUGGAACCGGAAGCGGAUCAGCAAUUGCGUGGCCGACUACUCCGUGCUGUACAACUCCGCCAGCUUC

VacPrv: AGCACCUUCAAGUGCUACGGCGUGUCCCCUACCAAGCUGAACGACCUGUGCUUCACAAACGUGUACGCCGACAGCUUCGUGAUCCGGGGAGAUGAAGUGCGG
VacPap: AGCACCUUCAAGUGCUACGGCGUGUCCCCUACCAAGCUGAACGACCUGUGCUUCACAAACGUGUACGCCGACAGCUUCGUGAUCCGGGGAGAUGAAGUGCGG

VacPrv: CAGAUUGCCCCUGGACAGACAGGCAACAUCGCCGACUACAACUACAAGCUGCCCGACGACUUCACCGGCUGUGUGAUUGCCUGGAACAGCAACAACCUGGAC
VacPap: CAGAUUGCCCCUGGACAGACAGGCAACAUCGCCGACUACAACUACAAGCUGCCCGACGACUUCACCGGCUGUGUGAUUGCCUGGAACAGCAACAACCUGGAC

VacPrv: UCCAAAGUCGGCGGCAACUACAAUUACCUGUACCGGCUGUUCCGGAAGUCCAAUCUGAAGCCCUUCGAGCGGGACAUCUCCACCGAGAUCUAUCAGGCCGGC
VacPap: UCCAAAGUCGGCGGCAACUACAAUUACCUGUACCGGCUGUUCCGGAAGUCCAAUCUGAAGCCCUUCGAGCGGGACAUCUCCACCGAGAUCUAUCAGGCCGGC

VacPrv: AGCACCCCUUGUAACGGCGUGAAGGGCUUCAACUGCUACUUCCCACUGCAGUCCUACGGCUUUCAGCCCACAUAUGGCGUGGGCUAUCAGCCCUACAGAGUG
VacPap: AGCACCCCUUGUAACGGCGUGAAGGGCUUCAACUGCUACUUCCCACUGCAGUCCUACGGCUUUCAGCCCACAUAUGGCGUGGGCUAUCAGCCCUACAGAGUG

VacPrv: GUGGUGCUGAGCUUCGAACUGCUGCAUGCCCCUGCCACAGUGUGCGGCCCUAAGAAAAGCACCAAUCUCGUGAAGAACAAAUGCGUGAACUUCAACUUCAAC
VacPap: GUGGUGCUGAGCUUCGAACUGCUGCAUGCCCCUGCCACAGUGUGCGGCCCUAAGAAAAGCACCAAUCUCGUGAAGAACAAAUGCGUGAACUUCAACUUCAAC

VacPrv: GGCCUGACCGGCACCGGCGUGCUGACAGAGAGCAACAAGAAGUUCCUGCCAUUCCAGCAGUUUGGCCGGGAUAUCGCCGAUACCACAGACGCCGUUAGAGAU
VacPap: GGCCUGACCGGCACCGGCGUGCUGACAGAGAGCAACAAGAAGUUCCUGCCAUUCCAGCAGUUUGGCCGGGAUAUCGCCGAUACCACAGACGCCGUUAGAGAU

VacPrv: CCCCAGACACUGGAAAUCCUGGACAUCACCCCUUGCAGCUUCGGCGGAGUGUCUGUGAUCACCCCUGGCACCAACACCAGCAAUCAGGUGGCAGUGCUGUAC
VacPap: CCCCAGACACUGGAAAUCCUGGACAUCACCCCUUGCAGCUUCGGCGGAGUGUCUGUGAUCACCCCUGGCACCAACACCAGCAAUCAGGUGGCAGUGCUGUAC

VacPrv: CAGGGCGUGAACUGUACCGAAGUGCCCGUGGCCAUUCACGCCGAUCAGCUGACACCUACAUGGCGGGUGUACUCCACCGGCAGCAAUGUGUUUCAGACCAGA
VacPap: CAGGGAGUGAACUGUACCGAAGUGCCCGUGGCCAUUCACGCCGAUCAGCUGACACCUACAUGGCGGGUGUACUCCACCGGCAGCAAUGUGUUUCAGACCAGA
              ^
VacPrv: GCCGGCUGUCUGAUCGGAGCCGAGCACGUGAACAAUAGCUACGAGUGCGACAUCCCCAUCGGCGCUGGAAUCUGCGCCAGCUACCAGACACAGACAAACAGC
VacPap: GCCGGCUGUCUGAUCGGAGCCGAGCACGUGAACAAUAGCUACGAGUGCGACAUCCCCAUCGGCGCUGGAAUCUGCGCCAGCUACCAGACACAGACAAACAGC

VacPrv: CCUCGGAGAGCCAGAAGCGUGGCCAGCCAGAGCAUCAUUGCCUACACAAUGUCUCUGGGCGUGGAGAACAGCGUGGCCUACUCCAACAACUCUAUCGCUAUC
VacPap: CCUCGGAGAGCCAGAAGCGUGGCCAGCCAGAGCAUCAUUGCCUACACAAUGUCUCUGGGCGUGGAGAACAGCGUGGCCUACUCCAACAACUCUAUCGCUAUC

VacPrv: CCCACCAACUUCACCAUCAGCGUGACCACAGAGAUCCUGCCUGUGUCCAUGACCAAGACCAGCGUGGACUGCACCAUGUACAUCUGCGGCGAUUCCACCGAG
VacPap: CCCACCAACUUCACCAUCAGCGUGACCACAGAGAUCCUGCCUGUGUCCAUGACCAAGACCAGCGUGGACUGCACCAUGUACAUCUGCGGCGAUUCCACCGAG

VacPrv: UGCUCCAACCUGCUGCUGCAGUACGGCAGCUUCUGCACCCAGCUGAAUAGAGCCCUGACAGGGAUCGCCGUGGAACAGGACAAGAACACCCAAGAGGUGUUC
VacPap: UGCUCCAACCUGCUGCUGCAGUACGGCAGCUUCUGCACCCAGCUGAAUAGAGCCCUGACAGGGAUCGCCGUGGAACAGGACAAGAACACCCAAGAGGUGUUC

VacPrv: GCCCAAGUGAAGCAGAUCUACAAGACCCCUCCUAUCAAGGACUUCGGCGGCUUCAAUUUCAGCCAGAUUCUGCCCGAUCCUAGCAAGCCCAGCAAGCGGAGC
VacPap: GCCCAAGUGAAGCAGAUCUACAAGACCCCUCCUAUCAAGGACUUCGGCGGCUUCAAUUUCAGCCAGAUUCUGCCCGAUCCUAGCAAGCCCAGCAAGCGGAGC

VacPrv: UUCAUCGAGGACCUGCUGUUCAACAAAGUGACACUGGCCGACGCCGGCUUCAUCAAGCAGUAUGGCGAUUGUCUGGGCGACAUUGCCGCCAGGGAUCUGAUU
VacPap: UUCAUCGAGGACCUGCUGUUCAACAAAGUGACACUGGCCGACGCCGGCUUCAUCAAGCAGUAUGGCGAUUGUCUGGGCGACAUUGCCGCCAGGGAUCUGAUU

VacPrv: UGCGCCCAGAAGUUUAACGGACUGACAGUGCUGCCUCCUCUGCUGACCGAUGAGAUGAUCGCCCAGUACACAUCUGCCCUGCUGGCCGGCACAAUCACAAGC
VacPap: UGCGCCCAGAAGUUUAACGGACUGACAGUGCUGCCUCCUCUGCUGACCGAUGAGAUGAUCGCCCAGUACACAUCUGCCCUGCUGGCCGGCACAAUCACAAGC

VacPrv: GGCUGGACAUUUGGAGCAGGCGCCGCUCUGCAGAUCCCCUUUGCUAUGCAGAUGGCCUACCGGUUCAACGGCAUCGGAGUGACCCAGAAUGUGCUGUACGAG
VacPap: GGCUGGACAUUUGGAGCAGGCGCCGCUCUGCAGAUCCCCUUUGCUAUGCAGAUGGCCUACCGGUUCAACGGCAUCGGAGUGACCCAGAAUGUGCUGUACGAG

VacPrv: AACCAGAAGCUGAUCGCCAACCAGUUCAACAGCGCCAUCGGCAAGAUCCAGGACAGCCUGAGCAGCACAGCAAGCGCCCUGGGAAAGCUGCAGGACGUGGUC
VacPap: AACCAGAAGCUGAUCGCCAACCAGUUCAACAGCGCCAUCGGCAAGAUCCAGGACAGCCUGAGCAGCACAGCAAGCGCCCUGGGAAAGCUGCAGGACGUGGUC
```

```
VacPrv: AACCAGAAUGCCCAGGCACUGAACACCCUGGUCAAGCAGCUGUCCUCCAACUUCGGCGCCAUCAGCUCUGUGCUGAACGAUAUCCUGAGCAGACUGGACCCU
VacPap: AACCAGAAUGCCCAGGCACUGAACACCCUGGUCAAGCAGCUGUCCUCCAACUUCGGCGCCAUCAGCUCUGUGCUGAACGAUAUCCUGAGCAGACUGGACCCU

VacPrv: CCUGAGGCCGAGGUGCAGAUCGACAGACUGAUCACAGGCAGACUGCAGAGCCUCCAGACAUACGUGACCCAGCAGCUGAUCAGAGCCGCCGAGAUUAGAGCC
VacPap: CCUGAGGCCGAGGUGCAGAUCGACAGACUGAUCACAGGCAGACUGCAGAGCCUCCAGACAUACGUGACCCAGCAGCUGAUCAGAGCCGCCGAGAUUAGAGCC

VacPrv: UCUGCCAAUCUGGCCGCCACCAAGAUGUCUGAGUGUGUGCUGGGCCAGAGCAAGAGAGUGGACUUUUGCGGCAAGGGCUACCACCUGAUGAGCUUCCCUCAG
VacPap: UCUGCCAAUCUGGCCGCCACCAAGAUGUCUGAGUGUGUGCUGGGCCAGAGCAAGAGAGUGGACUUUUGCGGCAAGGGCUACCACCUGAUGAGCUUCCCUCAG

VacPrv: UCUGCCCCUCACGGCGUGGUGUUUCUGCACGUGACAUAUGUGCCCGCUCAAGAGAAGAAUUUCACCACCGCUCCAGCCAUCUGCCACGACGGCAAAGCCCAC
VacPap: UCUGCCCCUCACGGCGUGGUGUUUCUGCACGUGACAUAUGUGCCCGCUCAAGAGAAGAAUUUCACCACCGCUCCAGCCAUCUGCCACGACGGCAAAGCCCAC

VacPrv: UUUCCUAGAGAAGGCGUGUUCGUGUCCAACGGCACCCAUUGGUUCGUGACACAGCGGAACUUCUACGAGCCCCAGAUCAUCACCACCGACAACACCUUCGUG
VacPap: UUUCCUAGAGAAGGCGUGUUCGUGUCCAACGGCACCCAUUGGUUCGUGACACAGCGGAACUUCUACGAGCCCCAGAUCAUCACCACCGACAACACCUUCGUG

VacPrv: UCUGGCAACUGCGACGUCGUGAUCGGCAUUGUGAACAAUACCGUGUACGACCCUCUGCAGCCCGAGCUGGACAGCUUCAAAGAGGAACUGGACAAGUACUUU
VacPap: UCUGGCAACUGCGACGUCGUGAUCGGCAUUGUGAACAAUACCGUGUACGACCCUCUGCAGCCCGAGCUGGACAGCUUCAAAGAGGAACUGGACAAGUACUUU

VacPrv: AAGAACCACACAAGCCCCGACGUGGACCUGGGCGAUAUCAGCGGAAUCAAUGCCAGCGUCGUGAACAUCCAGAAAGAGAUCGACCGGCUGAACGAGGUGGCC
VacPap: AAGAACCACACAAGCCCCGACGUGGACCUGGGCGAUAUCAGCGGAAUCAAUGCCAGCGUCGUGAACAUCCAGAAAGAGAUCGACCGGCUGAACGAGGUGGCC

VacPrv: AAGAAUCUGAACGAGAGCCUGAUCGACCUGCAAGAACUGGGGAAGUACGAGCAGUACAUCAAGUGGCCCUGGUACAUCUGGCUGGGCUUUAUCGCCGGACUG
VacPap: AAGAAUCUGAACGAGAGCCUGAUCGACCUGCAAGAACUGGGGAAGUACGAGCAGUACAUCAAGUGGCCCUGGUACAUCUGGCUGGGCUUUAUCGCCGGACUG

VacPrv: AUUGCCAUCGUGAUGGUCACAAUCAUGCUGUGUUGCAUGACCAGCUGCUGUAGCUGCCUGAAGGGCUGUUGUAGCUGUGGCAGCUGCUGCAAGUUCGACGAG
VacPap: AUUGCCAUCGUGAUGGUCACAAUCAUGCUGUGUUGCAUGACCAGCUGCUGUAGCUGCCUGAAGGGCUGUUGUAGCUGUGGCAGCUGCUGCAAGUUCGACGAG

VacPrv: GACGAUUCUGAGCCCGUGCUGAAGGGCGUGAAACUGCACUACACA
VacPap: GACGAUUCUGAGCCCGUGCUGAAGGGCGUGAAACUGCACUACACA
```

As we used the codons of the Pfizer vacine to encode all amino acids that did not differ vs the Wuhan variant, the previous 1 nucleotide difference we saw between the vaccine bases dissappeard, and instead we have the following differences:

1. The **LAL** sequence that is not present in the old vaccine we developed,
2. The **GGU** to **GGA** encoding that we did on purpose on the new vaccine (to conform to the GC level of the original Wuhan vaccine), but ommitted from the previous vaccine we created.

Other than that the two vaccines are completely identical (as in addition to using the same codons for the amino acids - based on the Pfizer vaccine, we also used the same LUT we created).

This means that after we will add the end codon, and the 5' and 3' UTRs, these will remain the only differences between the two vaccines.

Yaniv Erlich

```
# Add missing parts to the SA vaccine (end codon will be taken together with the 3' UTR from the Pfizer vaccine):
sa_vac_final_paper = f"{vaccines_mRNA_seq[company_names[0]][:vaccines_data[company_names[0]]['Start Codon Index']]}{sa_vaccine_
```

```
step = 136

for start_idx in range(0, len(sa_vac_final_paper), step):
    print(sa_vac_final_paper[start_idx:start_idx + step])
```

```
GAGAAUAAACUAGUAUUCUUCUGGUCCCCACAGACUCAGAGAGAACCCGCCACCAUGUUCGUGUUCCUGGUGCUGCUGCCUCUGGUGUCCAGCCAGUGUGUGAACCUGACCACCAGAACACAGCUGCCU
ACACCAACAGCUUUACCAGAGGCGUGUACUACCCCGACAAGGUGUUCAGAUCCAGCGUGCUGCACUCUACCCAGGACCUGUUCCUGCCUUUCUUCAGCAACGUGACCUGGUUCCACGCCAUCCACGUGU
CAAUGGCACCAAGAGAUUCGCUAACCCCGUGCUGCCCUUCAACGACGGGGUGUACUUUGCCAGCACCGAGAAGUCCAACAUCAUCAGAGGCUGGAUCUUCGGCACCACACUGGACAGCAAGACCCAGAC
AUCGUGAACAACGCCACCAACGUGGUCAUCAAAGUGUGCGAGUUCCAGUUCUGCAACGACCCCUUCCUGGGCGUCUACUACCACAAGAACAACAAGAGCUGGAUGGAAAGCGAGUUCCGGGUGUACAGC
ACAACUGCACCUUCGAGUACGUGUCCCAGCCUUUCCUGAUGGACCUGGAAGGCAAGCAGGGCAACUUCAAGAACCUGCGCGAGUUCGUGUUUAAGAACAUCGACGGCUACUUCAAGAUCUACAGCAAGC
UAUCAACCUCGUGCGGGGACUGCCUCAGGGCUUCUCUGCUCUGGAACCCCUGGUGGAUCUGCCCAUCGGCAUCAACAUCACCCGGUUUCAGACACUGCUGGCCCUGCACAGAAGCUACCUGACACCUGC
AGCAGCGGAUGGACAGCUGGUGCCGCCGCUUACUAUGUGGGCUACCUGCAGCCUAGAACCUUCCUGCUGAAGUACAACGAGAACGGCACCAUCACCGACGCCGUGGAUUGUGCUCUGGAUCCUCUGAGC
AGUGCACCCUGAAGUCCUUCACCGUGGAAAAGGGCAUCUACCAGACCAGCAACUUCCGGGUGCAGCCCACCGAAUCCAUCGUGCGGUUCCCCAAUAUCACCAAUCUGUGCCCCUUCGGCGAGGUGUUCA
CAGAUUCGCCUCUGUGUACGCCUGGAACCGGAAGCGGAUCAGCAAUUGCGUGGCCGACUACUCCGUGCUGUACAACUCCGCCAGCUUCAGCACCUUCAAGUGCUACGGCGUGUCCCCUACCAAGCUGAA
UGCUUCACAAACGUGUACGCCGACAGCUUCGUGAUCCGGGGAGAUGAAGUGCGGCAGAUUGCCCCUGGACAGACAGGCAACAUCGCCGACUACAACUACAAGCUGCCCGACGACUUCACCGGCUGUGUC
GGAACAGCAACAACCUGGACUCCAAAGUCGGCGGCAACUACAAUUACCUGUACCGGCUGUUCCGGAAGUCCAAUCUGAAGCCCUUCGAGCGGGACAUCUCCACCGAGAUCUAUCAGGCCGGCAGCACCC
CGGCGUGAAGGGCUUCAACUGCUACUUCCCACUGCAGUCCUACGGCUUUCAGCCCACAUAUGGCGUGGGCUAUCAGCCCUACAGAGUGGUGGUGCUGAGCUUCGAACUGCUGCAUGCCCCUGCCACAGU
CCUAAGAAAAGCACCAAUCUCGUGAAGAACAAAUGCGUGAACUUCAACUUCAACGGCCUGACCGGCACCGGCGUGCUGACAGAGAGCAACAAGAAGUUCCUGCCAUUCCAGCAGUUUGGCCGGGAUAUC
CCACAGACGCCGUUAGAGAUCCCCAGACACUGGAAAUCCUGGACAUCACCCCUUGCAGCUUCGGCGGAGUGUCUGUGAUCACCCCUGGCACCAACACCAGCAAUCAGGUGGCAGUGCUGUACCAGGGAC
UACCGAAGUGCCCGUGGCCAUUCACGCCGAUCAGCUGACACCUACAUGGCGGGUGUACUCCACCGGCAGCAAUGUGUUUCAGACCAGAGCCGGCUGUCUGAUCGGAGCCGAGCACGUGAACAAUAGCUA
GACAUCCCCAUCGGCGCUGGAAUCUGCGCCAGCUACCAGACACAGACAAACAGCCCUCGGAGAGCCAGAAGCGUGGCCAGCCAGAGCAUCAUUGCCUACACAAUGUCUCUGGGCGUGGAGAACAGCGUC
CCAACAACUCUAUCGCUAUCCCCACCAACUUCACCAUCAGCGUGACCACAGAGAUCCUGCCUGUGUCCAUGACCAAGACCAGCGUGGACUGCACCAUGUACAUCUGCGGCGAUUCCACCGAGUGCUCCA
GCUGCAGUACGGCAGCUUCUGCACCCAGCUGAAUAGAGCCCUGACAGGGAUCGCCGUGGAACAGGACAAGAACACCCAAGAGGUGUUCGCCCAAGUGAAGCAGAUCUACAAGACCCCUCCUAUCAAGGA
GGCUUCAAUUUCAGCCAGAUUCUGCCCGAUCCUAGCAAGCCCAGCAAGCGGAGCUUCAUCGAGGACCUGCUGUUCAACAAAGUGACACUGGCCGACGCCGGCUUCAUCAAGCAGUAUGGCGAUUGUCUC
UUGCCGCCAGGGAUCUGAUUUGCGCCCAGAAGUUUAACGGACUGACAGUGCUGCCUCCUCUGCUGACCGAUGAGAUGAUCGCCCAGUACACAUCUGCCCUGCUGGCCGGCACAAUCACAAGCGGCUGGA
AGCAGGCGCCGCUCUGCAGAUCCCCUUUGCUAUGCAGAUGGCCUACCGGUUCAACGGCAUCGGAGUGACCCAGAAUGUGCUGUACGAGAACCAGAAGCUGAUCGCCAACCAGUUCAACAGCGCCAUCGC
CAGGACAGCCUGAGCAGCACAGCAAGCGCCCUGGGAAAGCUGCAGGACGUGGUCAACCAGAAUGCCCAGGCACUGAACACCCUGGUCAAGCAGCUGUCCUCCAACUUCGGCGCCAUCAGCUCUGUGCUC
UCCUGAGCAGACUGGACCCUCCUGAGGCCGAGGUGCAGAUCGACAGACUGAUCACAGGCAGACUGCAGAGCCUCCAGACAUACGUGACCCAGCAGCUGAUCAGAGCCGCCGAGAUUAGAGCCUCUGCCA
CGCCACCAAGAUGUCUGAGUGUGUGCUGGGCCAGAGCAAGAGAGUGGACUUUUGCGGCAAGGGCUACCACCUGAUGAGCUUCCCUCAGUCUGCCCCUCACGGCGUGGUGUUUCUGCACGUGACAUAUGU
CAAGAGAAGAAUUUCACCACCGCUCCAGCCAUCUGCCACGACGGCAAAGCCCACUUUCCUAGAGAAGGCGUGUUCGUGUCCAACGGCACCCAUUGGUUCGUGACACAGCGGAACUUCUACGAGCCCCAC
CCACCGACAACACCUUCGUGUCUGGCAACUGCGACGUCGUGAUCGGCAUUGUGAACAAUACCGUGUACGACCCUCUGCAGCCCGAGCUGGACAGCUUCAAAGAGGAACUGGACAAGUACUUUAAGAACC
CCCCGACGUGGACCUGGGCGAUAUCAGCGGAAUCAAUGCCAGCGUCGUGAACAUCCAGAAAGAGAUCGACCGGCUGAACGAGGUGGCCAAGAAUCUGAACGAGAGCCUGAUCGACCUGCAAGAACUGGG
GAGCAGUACAUCAAGUGGCCCUGGUACAUCUGGCUGGGCUUUAUCGCCGGACUGAUUGCCAUCGUGAUGGUCACAAUCAUGCUGUGUUGCAUGACCAGCUGCUGUAGCUGCCUGAAGGGCUGUUGUAGC
GCUGCUGCAAGUUCGACGAGGACGAUUCUGAGCCCGUGCUGAAGGGCGUGAAACUGCACUACACAUGAUGACUCGAGCUGGUACUGCAUGCACGCAAUGCUAGCUGCCCCUUUCCCGUCCUGGGUACCC
CCCCCGACCUCGGGUCCCAGGUAUGCUCCCACCUCCACCUGCCCCACUCACCACCUCUGCUAGUUCCAGACACCUCCCAAGCACGCAGCAAUGCAGCUCAAAACGCUUAGCCUAGCCACACCCCCACGC
CAGUGAUUAACCUUUAGCAAUAAACGAAAGUUUAACUAAGCUAUACUAACCCCAGGGUUGGUCAAUUUCGUGCCAGCCACACCCUGGAGCUAGCA
```

Conclusions for this section

Despite the fact that we based this vaccine on completely different samples than the previous one we created (all previous samples were shorter by 3 amino acids), we still got exactly the same vaccine for the SA variant, with the only differences being either the extra/missing **LAL** amino sequence, or the ones introduced by us.

The meaning of this is that all the proteins we analyzed in all 3 phases above (looking only at sequences without sequencing errors, looking at all sequences as long as these have the missing **LAL** amino sequence, or based on the

11:22 PM, May 12

Correct! amazing job.

mutations described in the paper) are identical (down to the **LAL** sequence and sequencing ambiguities), and that the only difference between them is on the nucleotide level (we do not notice this in the comparison above, as the codons used are taken from the Pfizer vaccine).

Conclusions

1. We started off by familiarizing ourselves with the Wuhan variant of the Covid19 virus and the Pfizer vaccine for it, on both the nucleotide and amino acids levels. As part of this process we found:
    1. That the vaccine's spike protein differs from the virus' on only 2 amino acids, that are in charge of making the physical form of the protein less prone to change before it contacts a host's cell.
    2. That the GC content of the vaccine is much higher than that of the virus, in order to weaken the virus.
    3. That a LUT of sorts exists between the codons of the vaccine and those of the virus, which explains the difference in the GC content between them.
2. We continued to trying to develop a vaccine for the SA variant. This included the following steps:
    1. First we tried to identify samples belonging to the SA variant (out of all the complete human samples available on NCBI). We did this by comparing known mutation sites and their surroundings.
    2. We then tried to cull these samples by searching for unmatching dates or countries of collection, but did not find any samples to cull. We did however cull samples which could not be properly translated from nucleotides to amino acids (which we added to the analysis separately, and eventually concluded would not affect the result).
    3. Later we tried to synthesize the SA variant virus' spike protein by comparing the proteins of the the different samples. This led us to cull another sample based on its similarity to the Wuhan variant's virus and vaccine spike protein.
    4. This allowed us to synthesize the SA variant virus's spike protein by using all the samples' spike protein (and using majority vote first on the amino then on the nucleotide levels in cases of disagreement).
    5. After that, we created a base for our new SA variant vaccine (the spike protein of the vaccine) by using the synthesised virus spike protein, and the aformentioned LUT between virus and vaccine codons.
    6. We finished by adding the end codon, and the 3' and 5' UTRs from the Pfizer vaccine for the Wuhan variant, and by that finished creating our vaccine.

7. In addition, we compared our initial findings - a vaccine based on the "shorter" **LAL** missing spike protein, to using samples that have the longer **LAL** included spike protein (based on a paper), and arrived at the conclusion that other than the **LAL** sequence itself, and 2 other nucleotide changes (that **we** introduced to keep the GC content similar to the original Pfizer vaccine), the vaccines would have turned out identical.